gLOP: the global and Local Penalty for Capturing Predictive Heterogeneity

Rhiannon V. Rose rrose24@uwo.ca
Department of Epidemiology & Biostatistics
Western University
London, ON, Canada

Daniel J. Lizotte dlizotte@uwo.ca
Department of Computer Science, Department of Epidemiology & Biostatistics
Western University
London, ON, Canada

Abstract

When faced with a supervised learning problem, we hope to have rich enough data to build a model that predicts future instances well. However, in practice, problems can exhibit *predictive heterogeneity*: most instances might be relatively easy to predict, while others might be *predictive outliers* for which a model trained on the entire dataset does not perform well. Identifying these can help focus future data collection. We present gLOP, the global and Local Penalty, a framework for capturing predictive heterogeneity and identifying predictive outliers. gLOP is based on penalized regression for multitask learning, which improves learning by leveraging training signal information from related tasks. We give two optimization algorithms for gLOP, one space-efficient, and another giving the full regularization path. We also characterize uniqueness in terms of the data and tuning parameters, and present empirical results on synthetic data and on two health research problems.

1. Introduction

We are motivated by prediction problems in healthcare where we have data about a population of patients, but only limited data about each individual patient. As an example, we will present a problem where the goal is to use a self-reported scale of depressive symptoms to predict a clinician-rated scale which has been deemed more suitable for decision-making. We have data on hundreds of different patients, but no more than 15 paired observations per patient. We would like to determine whether a single model relating self-reported scores to clinician-rated scores predicts all patients adequately, or if there are patients for whom such a model needs tailoring to work well for them. We also present another problem where the goal is to predict Parkinson's disease symptom progression from speech waveform features. Both of these problems have potential applications in telehealth. In studies of depression, although self-reported symptom scores are often collected, clinician-rated scores are commonly used to support decisions about altering a patient's treatment plan (Rush et al., 2004). For patients who may not have regular access to a trained assessor, the combination of a self-reported score and a prediction of the clinician-rated score could enable more timely revision of their treatment plan. The Parkinson's disease study had a similar intent; effective predictive models would help remote patients better track their symptoms and inform any revisions to their treatment plan.

In both of these problems (and in many others) we expect a degree of *predictive heterogeneity.* That is, we expect that for each patient, we could build an effective predictive model given enough patient-specific examples, but: 1) we do not have access to enough data per patient and 2) we suspect that individual heterogeneity will preclude building a "global model" that works well for all patients. We expect that some patients will be *predictive outliers* whose models differ substantially from the norm. While this cannot directly help us build better models for new patients, if we find evidence that predictive outliers exist, then we may be able to improve prediction by gathering more features on the patients we already have in order to help distinguish such outliers *a priori* in future models, and thus improve predictive power. If we see no evidence for predictive outliers, a better strategy might be to gather more training examples rather than more features on the existing examples.

Our contributions are as follows. We present the global and Local Penalty (gLOP) model, which learns a predictive model with a *global* component that applies to all patients and a *local* component that captures individual variation, while performing simultaneous feature selection for both. We describe in detail how it is related to previous approaches, and we present two optimization techniques for gLOP, one that is very space-efficient and one that provides the entire regularization path. We characterize the conditions under which the gLOP estimate is unique. We provide empirical evidence that gLOP has better in-population predictive performance than previous approaches. Finally, we show how gLOP can be used to detect predictive outliers by applying it to two health research problems.

2. Background

Lowercase letters (e.g. c, α) denote scalars, bold lowercase letters denote vectors (e.g. $\boldsymbol{y}$), and uppercase letters denote matrices (e.g. L). Superscripts index elements of a list (e.g. L^k is the kth matrix in a list) and subscripts index matrix columns (e.g. L^k_j is the jth column of the kth matrix) or vector elements (e.g. α_j is the jth element of the vector $\boldsymbol{\alpha}$.)

2.1 Penalization and the Lasso

Both gLOP and its related methods are based on feature selection through *penalization* or *regularization.* Given an $n \times p$ design matrix of predictor variables X, and a binary $n \times 1$ response variable $\boldsymbol{y}$, the general form of a penalized regression problem is $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \mathcal{L}(X, \boldsymbol{y}) + \mathcal{P}(\boldsymbol{\beta})$, where $\mathcal{L}$ is a loss function measuring how well the model $\boldsymbol{\beta}$ fits the data, and $\mathcal{P}$ is a penalty term on the complexity of $\boldsymbol{\beta}$. The *Least Absolute Shrinkage and Selection Operator* (lasso) (Tibshirani, 1996) is one such method, whose parameter estimates are given by $\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda) = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$, where λ controls the amount of shrinkage of the estimates. When $\lambda = 0$, the model is unpenalized with all features present; higher values of λ will cause more coefficients of $\boldsymbol{\beta}$ to be shrunk to 0, giving a sparser model that includes only the most relevant features.

2.2 Penalization Methods for Multi-task Learning

Multi-task learning methods (Caruana, 1997) were originally described as learning from data about a large number of related "tasks" when there is only limited data about each individual task. This framework has a clear connection to the problem we face with limited

patient data. We make one terminological change in this work: We use the word "patient" instead of "task," to clarify that there is only one predictive "task," e.g., we are always trying to predict a symptom score. The data are divided into different patients from which we have observations relevant to performing the task.

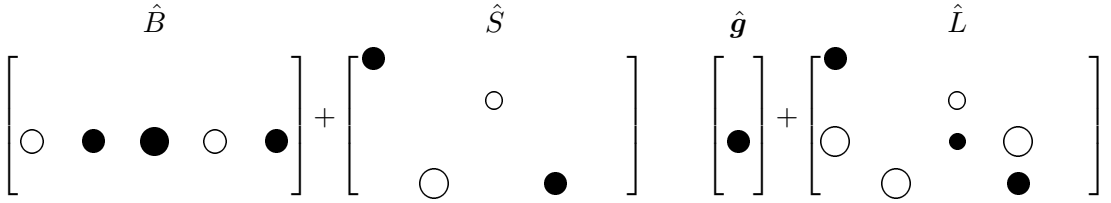
The *Composite Absolute Penalties* (CAP) family of models, of which the lasso is a special case, is used in various cases of hierarchical and group-based feature selection. Intuitively, CAP penalties work by taking the norms of vectors that contain coefficients of different groups of variables, and then penalizing the norm of the vector containing each of the group norms. Construction of a general CAP works as follows. For each of κ groups of coefficients, we create sub-vectors of coefficients denoted $\boldsymbol{\beta}^k$. We take the norms $\nu_k = \|\boldsymbol{\beta}^k\|_{\gamma_k}$ of each of these sub-vectors, and place them in a κ-dimensional vector $\boldsymbol{\nu} = (\nu_1, ..., \nu_\kappa)$. The CAP penalty is given by $\|\boldsymbol{\nu}\|_{\gamma_0}^{\gamma_0} = \sum_k |\nu_k|^{\gamma_0}$. Using a least-squares loss, the CAP estimate as a function of λ is given by $\hat{\boldsymbol{\beta}}^{\text{CAP}}(\lambda, \boldsymbol{\gamma}) = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|y - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\nu}\|_{\gamma_0}^{\gamma_0}$. To induce sparsity across groups, we would select $\gamma_0 = 1$. CAPs are convex provided that the norms used are also convex, making global optimization feasible (Zhao et al., 2009).

In order to leverage data collected from different (but similar) tasks (we would say patients), Jalali et al. (2010) developed a CAP-based method used for what they term *dirty data*: data containing features that are not relevant to all tasks. They outlined a *dirty model* that can leverage similarity between tasks by identifying features that are globally relevant, but that can also allow for the inclusion of features that are only relevant in some tasks. This corresponds to the situation in which most patients may be well-served by the same predictive model, but some patients require different models. Let n_k and p denote the number of examples and features, respectively, *per patient.* There are κ patients. The dirty model is parameterized by a "global" matrix of parameters B and a "local" matrix of parameters S, both of which are $p \times \kappa$. The feature matrix for patient k is denoted by X^k, and the targets are denoted $\boldsymbol{y}^k$. The optimization problem for the dirty model is[1]

$$\underset{B,S}{\text{argmin}} \sum_{k=1}^{\kappa} \|\boldsymbol{y}^k - X^k(B_k + S_k)\|_2^2 + \lambda_B\|B\|_{1,\infty} + \lambda_S\|S\|_{1,1}. \tag{1}$$

The learned parameters for patient k are then given by $\boldsymbol{\beta}^k = \hat{B}_k + \hat{S}_k$, where $\hat{B}$ and $\hat{S}$ are solutions of (1). The dirty model applies an $\ell_{1,\infty}$ norm penalty to parameter matrix B, which is given by $\|B\|_{1,\infty} = \sum_j \|(B^\mathsf{T})_j\|_\infty$. Hence, this is a CAP problem that uses the ℓ_1 and ℓ_∞ norms to achieve group sparsity. The effect of this penalty is to induce entire rows of B to enter the model at the same time, that is, as soon as one element of B enters the model, all elements corresponding to the same feature in different patients may enter the model with no additional penalty as long as their absolute value stays less than or equal to that of the largest element. Thus parameters in B will "turn on" for all patients at once. Note however that there is no strict enforcement of equality across the rows of B, so while the selection is global, the actual parameter values are not. The secondary parameter matrix S is penalized using the $\ell_{1,1}$ norm, which is simply the sum of the absolute values of the elements in the matrix; this induces element-wise sparsity in $\hat{S}$ and allows individual patients to "turn on"

1. Here we give the objective function as defined in the code provided by Jalali et al.. It differs from the objective stated in their paper by a factor of $\frac{1}{2n}$ applied to the squared loss term.



(a) Example estimated dirty model structure with global parameters $\hat{B}$ and local parameters $\hat{S}$.

(b) Example estimated gLOP model structure with global parameters $\hat{g}$ and local parameters $\hat{L}$.

Figure 1: Schematic view of a dirty and a gLOP model that would make identical predictions. Columns of $\hat{B}$, $\hat{S}$, and $\hat{L}$ represent coefficients for each patient. Filled circles represent positive coefficients, empty circles represent negative. Sizes indicate magnitude.

additional features if necessary. A schematic view of a hypothetical learned $\hat{B}$ and $\hat{S}$ is shown in Figure 1a. Representing the parameters of the model as the matrices B and S allows the dirty model to capture some similarities between patients while still allowing for individual variation. However, the interpretation of the model is not straightforward because coefficients for a single feature in B, although considered the "global" part of the model, are not required to be the same for all patients. Furthermore, although the dirty model penalty is convex, it does not admit a straightforward regularization path algorithm.

3. gLOP

We now introduce our global and LOcal Penalty (gLOP) model. As for the "dirty model," we learn global and local parameters, but our decomposition uses a *vector* $\boldsymbol{g}$ (for "global") and a matrix L (for "local"). The $p \times 1$ vector $\boldsymbol{g}$ contains global coefficients that apply to all patients, and the columns of the $p \times \kappa$ matrix L contain local coefficients that apply only to their specific patients. This makes our model more easily interpretable, as the global effects are clearly distinguishable from individual effects and are enforced to be *the same across all patients.* Thus we apply a simpler ℓ_1 penalty to $\boldsymbol{g}$ instead of the $\ell_{1,\infty}$ norm used in the dirty model because we do not need to use the penalty to "push" the global parameters to be the same across patients. The gLOP optimization problem is given by

$$\underset{\boldsymbol{g},L}{\operatorname{argmin}} \sum_{k=1}^{\kappa} \tfrac{1}{2n_k} \|\boldsymbol{y}^k - X^k(\boldsymbol{g} + L_k)\|_2^2 + \lambda_{\boldsymbol{g}} \|\boldsymbol{g}\|_1 + \lambda_L \|L\|_{1,1}. \tag{2}$$

The learned parameters for patient k are then given by $\boldsymbol{\beta}^k = \boldsymbol{g} + \hat{L}_k$, where $\hat{\boldsymbol{g}}$ and $\hat{L}$ are given by (2). A diagram of an example $\hat{\boldsymbol{g}}$ and $\hat{L}$ is shown in Figure 1b.

Simplifying the matrix B into our vector $\boldsymbol{g}$ is advantageous as it reduces the number of parameters, which reduces the potential for overfitting compared to the dirty model. Additionally, because the global coefficients are identical across patients, the model is easily interpretable for all patients together and individual patients, increasing the utility of the model in scientific practice. This model formulation also offers computational advantages; we present a space-efficient block coordinate minimization with the lasso as a subroutine,

and we present a method that allows us to compute the full regularization path, which is not possible with the dirty model. Finally, the gLOP formulation allows us to establish deterministic conditions for uniqueness in terms of the data and the penalty parameters.

3.1 Block Coordinate Minimization

Problem (2) can be optimized using block coordinate minimization (Wright and Nocedal, 1999) using the standard lasso. A significant advantage of this method is its use of the lasso as a subroutine, for which fast implementations are commonly available. To solve (2) using the lasso, we decompose the optimization into separate problems for L and $\boldsymbol{g}$. If we fix L, the $\boldsymbol{g}$ that optimizes (2) is given by

$$\underset{\boldsymbol{g}}{\operatorname{argmin}} \tfrac{1}{2n_k}\|\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{g}\|_2^2 + \lambda_{\boldsymbol{g}}\|\boldsymbol{g}\|_1, \text{ where } \tilde{X} = \begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^\kappa \end{bmatrix} \tilde{y} = \begin{bmatrix} \boldsymbol{y}^1 - X^1 L_1 \\ \boldsymbol{y}^2 - X^2 L_2 \\ \vdots \\ \boldsymbol{y}^\kappa - X^\kappa L_\kappa \end{bmatrix}. \tag{3}$$

Problem (3) is a standard lasso problem. If we fix $\boldsymbol{g}$, the L that optimizes (2) is given by

$$\underset{L}{\operatorname{argmin}} \sum_{k=1}^{\kappa} \left[\tfrac{1}{2n_k}\|(\boldsymbol{y}^k - X^k\boldsymbol{g}) - X^k L_k\|_2^2 + \lambda_L\|L_k\|_1 \right]. \tag{4}$$

Note that each term in the sum in (4) involves only one column of L. Therefore we can optimize each column of L independently:

$$\underset{L^k}{\operatorname{argmin}} \tfrac{1}{2n_k}\|\check{\boldsymbol{y}}^k - X^k L_k\|_2^2 + \lambda_L\|L_k\|_1 \tag{5}$$

where X^k is the design matrix for patient k, L_k is the column of L for patient k and $\check{\boldsymbol{y}}^k = \boldsymbol{y}^k - X^k\boldsymbol{g}$, or $\boldsymbol{y}$ adjusted for the contribution of $\boldsymbol{g}$, where $\boldsymbol{y}^k$ is the vector of observations for patient k. Note that we are using squared loss in all of the above equations in our development, but any strictly convex loss function can be substituted, meaning that our approach applies to other generalized linear models as well. We can solve (2) by choosing a starting point and alternating solving problems (5) and (3) until some convergence criterion is met. Note that since (2) is convex, this procedure will converge to a global optimum, although that optimum may not be unique as we discuss below.

3.2 A Single-Lasso View of gLOP

The block coordinate minimization algorithm presented above works very well and is space-efficient, taking $\mathcal{O}(\sum_k n_k p)$ space, which allows gLOP to be applied to large datasets. However it only produces the gLOP estimate for a single pair of $\lambda_{\boldsymbol{g}}$ and λ_L. We now present an alternative formulation that allows us to recover the entire regularization path for gLOP, and also allows us to characterize the uniqueness of gLOP estimates. For simplicity, we assume the n_k are all equal. We define a $n \cdot \kappa \times p \cdot (\kappa + 1)$ block matrix with the first p columns containing the vertical concatenation of the design matrices for each patient,

horizontally concatenated with a block matrix with design matrices for each patient on the diagonal. We then define target and coefficient vectors $\boldsymbol{y}$ and $\boldsymbol{\beta}$ as follows[2]:

$$X = \begin{bmatrix} X^1 & X^1 & 0 & \cdots & 0 \\ X^2 & 0 & X^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X^\kappa & 0 & 0 & \cdots & X^\kappa \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}^1 \\ \boldsymbol{y}^2 \\ \vdots \\ \boldsymbol{y}^\kappa \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{g} \\ L_1 \\ L_2 \\ \vdots \\ L_\kappa \end{bmatrix}$$

We can then write the gLOP optimization as

$$\operatorname*{argmin}_{\boldsymbol{\beta}} \tfrac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{p\cdot(\kappa+1)} \lambda_i|\beta_i| \tag{6}$$

where we set $\lambda_i = \lambda_{\boldsymbol{g}}$ for $1 \le i \le p$ and $\lambda_i = \lambda_L$ for $p+1 \le i \le p\cdot(\kappa+1)$. Equation (6) is a lasso-like problem but with different regularization weights applied to different coefficients. Note that if we set $\xi_i = \lambda_i\beta_i$, we may re-write the problem equivalently as

$$\operatorname*{argmin}_{\boldsymbol{\xi}} \tfrac{1}{2n}\|\boldsymbol{y} - \bar{X}\boldsymbol{\xi}\|_2^2 + \|\boldsymbol{\xi}\|_1, \text{ where } \bar{X} = \begin{bmatrix} \frac{X^1}{\lambda_{\boldsymbol{g}}} & \frac{X^1}{\lambda_L} & 0 & \cdots & 0 \\ \frac{X^2}{\lambda_{\boldsymbol{g}}} & 0 & \frac{X^2}{\lambda_L} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{X^\kappa}{\lambda_{\boldsymbol{g}}} & 0 & 0 & \cdots & \frac{X^\kappa}{\lambda_L} \end{bmatrix}. \tag{7}$$

by absorbing the penalty parameters into the design matrix. This is a standard lasso problem with $\lambda = 1$, which we can solve and then recover the original estimates by defining $\hat{\beta}_i = \hat{\xi}_i/\lambda_i$ for $\lambda_i > 0$. This formulation gives us the ability to use properties of the lasso to develop optimization algorithms for and determine properties of the gLOP estimator. In the following sections, we develop an algorithm for the full regularization path of gLOP, and we give a complete characterization of the uniqueness of gLOP estimates.

3.3 The Full Regularization Path of gLOP

Least-Angle Regression (LARS) (Efron et al., 2004) was originally formulated as a non-greedy forward variable selection algorithm for least-squares regression. With a slight modification, the LARS algorithm produces the entire regularization path for the lasso; that is, it produces the lasso solutions for a given problem for all $\lambda \ge 0$ by expressing the solution path as a piecewise linear vector-valued function. In the following, we use "LARS" to mean the commonly-used lasso version of the LARS algorithm.

Using the formulation given in (7), we can apply the LARS algorithm to obtain a solution path for gLOP as we now show. By fixing "reference" values $\lambda_{\boldsymbol{g}}^*$ and λ_L^*, we apply LARS to the problem defined in (7), which under certain conditions (discussed further below) gives all solutions of the form $\hat{\boldsymbol{\xi}}(\lambda) = \operatorname{argmin}_{\boldsymbol{\xi}} \frac{1}{2n}\|\boldsymbol{y} - \bar{X}\boldsymbol{\xi}\|_2^2 + \lambda\|\boldsymbol{\xi}\|_1$ where $\hat{\boldsymbol{\xi}}(\lambda)$ gives the gLOP solution for $\lambda_{\boldsymbol{g}} = \lambda \cdot \lambda_{\boldsymbol{g}}^*$, and $\lambda_L = \lambda \cdot \lambda_L^*$. Thus we obtain all solutions corresponding

2. If the n_k are not all equal, we need additional normalization in the definitions of X, $\boldsymbol{y}$, and $\boldsymbol{\beta}$; we omit these details for clarity

to a fixed ratio $\lambda_{\boldsymbol{g}}/\lambda_L$, but whose "overall" amount of penalization varies according to λ. This may be interpreted as fixing the amount of preference for a more global versus a more local model, and varying the total number of parameters that are permitted to enter the model. The corresponding $\hat{\boldsymbol{\beta}}$ parameter is given by $\hat{\beta}_i(\lambda) = \hat{\xi}_i(\lambda)/\lambda^*_{\boldsymbol{g}}$ for $i \leq p$ (the global parameters) and $\hat{\beta}_i(\lambda) = \hat{\xi}_i(\lambda)/\lambda^*_L$ for $i > p$ (the local parameters.)

3.4 Tuning Parameters and Uniqueness

We now characterize the uniqueness of gLOP estimates in terms of the data and the tuning parameters. Proofs of the lemmas are given in Appendix A. We begin by extending work by Ryan Tibshirani (2013) to the case where not all lasso coefficients are penalized equally by using our modified design matrix $\bar{X}$. Our objective in (3.3) has subgradients evaluated at a point $(\boldsymbol{\xi}, \lambda)$ given by

$$-\bar{X}^\mathsf{T}(\boldsymbol{y} - \bar{X}\boldsymbol{\xi}) + \lambda\boldsymbol{\alpha}, \quad \text{s.t.} \quad \alpha_i \in \begin{cases} \{\text{sgn}(\xi_i)\} & \text{if } \xi_i \neq 0 \\ [-1, 1] & \text{if } \xi_i = 0. \end{cases} \tag{8}$$

First-order optimality conditions state that if a point $\hat{\boldsymbol{\xi}}$ is optimal, then there exist $\hat{\alpha}_i$ satisfying constraints in (8) for which the given expression is the zero vector. Because our objective is convex (but not strictly convex) these conditions are also sufficient for optimality, though there is no guarantee that the optimum is unique. For an optimal point $\hat{\boldsymbol{\xi}}$ with corresponding $\hat{\boldsymbol{\alpha}}$ satisfying first-order optimality conditions, define $\mathcal{A} = \{i \in \{1, ..., p\} : |\hat{\alpha}_i| = 1\}$ where $\bar{X}_i$ is the ith column of $\bar{X}$. This set contains all the indices of the non-zero coefficients, plus possibly indices of some of the zero coefficients. Let $\bar{X}_\mathcal{A}$ be the sub-matrix of $\bar{X}$ consisting of only the columns with indices in $\mathcal{A}$.

Lemma 1 *If $\bar{X}_\mathcal{A}$ has full rank then the lasso has a unique solution.*

Thus, one way of ensuring uniqueness of $\hat{\boldsymbol{\xi}}$ is to guarantee that $\bar{X}_\mathcal{A}$ is of full column rank. However, we do not know which columns of $\bar{X}$ will be in $\bar{X}_\mathcal{A}$ until we solve the optimization problem. We could ensure full rank of $\bar{X}_\mathcal{A}$ by guaranteeing that $\bar{X}$ is of full rank, but we do not wish to restrict our attention to full-rank design matrices; indeed one of the most useful problems for lasso-type methods is when $p > n$ and $\bar{X}$ obviously does not have full rank. Furthermore, for gLOP, the matrix $\bar{X}$ will *never* be full-rank. For example, suppose $p = 4$ and $\kappa = 3$; then by definition we have $\bar{X}_1 = \bar{X}_5 + \bar{X}_9 + \bar{X}_{13}$, establishing that $\bar{X}$ has linear dependence among its columns and is not full-rank.

We now show that a weaker condition – *affine independence with negation* – among columns of $\bar{X}$ is sufficient to ensure that $\bar{X}_\mathcal{A}$ is always of full rank.

Definition 2 (Affinely Independent with Negation (AIN)) *The columns of an $n \times p$ matrix X are* Affinely Independent with Negation *(AIN) if there* do not exist *signs s_i, weights w_i, and an index j such that $X_j = \sum_{\substack{i=1 \\ i \neq j}}^{p} w_i s_i X_i$, where $s_i \in \{-1, 1\}$, and $\sum_i w_i = 1$. (Standard affine independence is similar but does not allow for the s_i.)*

Lemma 3 *If the columns of $\bar{X}$ are AIN, then $\bar{X}_\mathcal{A}$ has full rank.*

The design matrix $\bar{X}$ is defined in terms of the "original" gLOP matrix X together with $\lambda_{\boldsymbol{g}}$ and λ_L. Each of these components influences whether or not the columns of $\bar{X}$ will be AIN; we summarize this in the following theorem.

Theorem 4 *If the the columns of each matrix X^k, $k = 1...\kappa$ are AIN, and if $\lambda_L > \lambda_{\boldsymbol{g}}$, then the columns of $\bar{X}$ are AIN and there is a unique gLOP solution.*

Proof First, we note that columns from different patient blocks are orthogonal; thus we cannot construct columns from one patient block using weighted sums of columns from other patient blocks. Furthermore, we assume that within each patient block we have AIN columns. (We will discuss this assumption more later.) Therefore, if there is affine dependence among columns, it must involve the patient blocks and the global block.

Each column in the global block of $\bar{X}$ can be written as a linear combination of columns from the patient blocks. In particular, for $i \leq p$, we have $X_i/\lambda_{\boldsymbol{g}} = \sum_{k=1}^{\kappa} \left(\frac{\lambda_L}{\lambda_{\boldsymbol{g}}}\right) X_{p\cdot k+i}/\lambda_L$. We could write this linear combination with weights $w_i = \frac{\lambda_L}{\lambda_{\boldsymbol{g}}}$ and signs $s_i = 1$, or we could also negate any number of the s_i along with their corresponding w_i to achieve the same result. If we negate k of the columns, the sum of the weights is given by $(\kappa - k)\frac{\lambda_L}{\lambda_{\boldsymbol{g}}} - k(\lambda_L/\lambda_{\boldsymbol{g}})$. Therefore to ensure the AIN property of $\bar{X}$, we may choose $\lambda_{\boldsymbol{g}}$ and λ_L such that $(\kappa - k)(\lambda_L/\lambda_{\boldsymbol{g}}) - k(\lambda_L/\lambda_{\boldsymbol{g}}) \neq 1$, or equivalently $(\kappa - 2k)\,\frac{\lambda_L}{\lambda_{\boldsymbol{g}}} \neq 1$ for all $0 \leq k \leq \kappa$. We can ensure that this holds by noting that if we choose $\lambda_L > \lambda_{\boldsymbol{g}}$, then the inequality holds because $(\kappa - 2k)$ is an integer and $\frac{\lambda_L}{\lambda_{\boldsymbol{g}}} > 1$, so their product cannot possibly equal 1. Note that if κ is even, then $\frac{\lambda_L}{\lambda_{\boldsymbol{g}}} > \frac{1}{2}$ is sufficient. ■

Theorem 4 characterizes uniqueness of gLOP in terms of the penalty parameters and the data matrices for each patient, X^1 through X^k. Tibshirani et al. (2013) note that a design matrix drawn from a continuous probability distribution on $\mathbb{R}^{np}$ has the AIN property with probability one (see Tibshirani et al., 2013, Lemma 4). Thus, for matrices of continuous feature values, the uniqueness of gLOP can be assured by an appropriate choice of $\lambda_{\boldsymbol{g}}$ and λ_L alone. Design matrices containing discrete entries require more careful analysis to ensure uniqueness. Another interesting consequence of this is that all of our lasso sub-problems for $\boldsymbol{g}$ and the L_k that are used as part of our block coordinate minimization algorithms have unique solutions under the same conditions on the X^k, even if the joint minimization does not have a unique solution. Finally, we note that Jalali et al. (2010) provide results addressing the uniqueness of the dirty model estimates asymptotically in n with high probability rather than with probability one as we have here.

3.5 Empirical Results: Synthetic Data

To evaluate the *in-population* predictive accuracy of gLOP and the dirty model in different contexts, we conduct four main experiments using different sizes of p and κ. By *in-population*, we mean accuracy on future data gathered on the same population of patients as were used for training. This is contrasted with *out-of-population* prediction, when a model is used to predict labels for instances of future (unseen) patients. For each experiment, we run 100 trials using data generated as described above and average the results.

Means and standard deviations of the MSE from each experiment are shown in Table 1; for comparison, we also include the MSE of the result achieved by the standard lasso, which essentially ignores the distinction between data from different patients. The detailed experimental setup is given in Appendix B. In all cases, the test error for gLOP is statistically significantly lower than the error for the dirty model and for the lasso. In the small-p settings, both gLOP and the dirty model do much better than the lasso because they allow different patients to have different coefficients, resulting in a much better model fit. For the large-p case, the error for gLOP decreases with κ, but the error for the dirty model remains the same. We attribute this to the increased number of parameters that the dirty model tries to learn; that is, $2 \cdot \kappa \cdot p$ parameters for the dirty model versus $p + \kappa \cdot p$ for gLOP, even though they have the same representational power.

We also conducted a simple experiment to illustrate how gLOP can identify predictive outliers. Consider a case where we have data from several patients and we want to construct a global model, but 20 percent of patients are predictive outliers in the sense that they have very different local intercepts from the remainder of the patients. The true model is given by $Y = 1 + X + cZ + \epsilon$, where $X \sim \mathcal{N}(0, 1)$, $Z \sim$ Bernoulli(0.2), ϵ $\mathcal{N}(0, 1)$ and $c = 10$. In this case, we can use gLOP to construct a model *using only* X *and* Y to identify predictive outliers by examining local coefficients. For our example, we generated synthetic data for $\kappa = 16$ patients, $n_k = 10$ observations per patient, and $p = 32$ features. Using only X and Y, gLOP correctly identified the 5 outliers (for whom $Z = 1$) by assigning them larger non-zero local intercepts. Adding Z into the model results in improved performance of the global model and no detection of predictive outliers. In a clinical setting, we envision that gLOP could be used to first identify predictive outliers, which would then direct the search for a new feature that could identify them and in turn improve the predictions of the global model.

4. Empirical Results: Health Research Data

We now present two examples of how gLOP can be used to identify predictive outliers and direct future data gathering to improve predictive performance. The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study was a multi-stage, multi-centre prospective randomized clinical trial assessing interventions for non-psychotic major depressive disorder (MDD) in the context of sequential alternatives for MDD treatment (Rush et al., 2004). In the study, proceeding to a new treatment was contingent on the clinician-rated Quick Inventory for Depressive Symptomatology (QIDS; Rush et al. (2003); lower scores are better). If a patient did not improve according to this scale, a new treatment was initiated. According to the protocol, self-reported QIDS was also recorded at each clinic visit. We used gLOP to predict clinician-rated QIDS from self-report QIDS using data from 1,368 patients over 15,593 visits, using patient age as an additional demographic variable in g only. λ values were chosen using the Bayesian Information Criterion (BIC) (Zou et al., 2007) as for some patients there were not enough observations to support cross-validation.

The global intercept and slope revealed that patients on average tended to rate their own symptoms lower than clinicians did. We found that 25 patients had a local intercept, 32 patients had a nonzero local slope coefficient, and 2 patients had both a local intercept and slope. Of the patients that had a local intercept, all but 3 had a positive coefficient

indicating that this group of patients tended to underrate their symptoms even more than the general population relative to the clinician-rated scores. Of the patients that had a nonzero local slope, all but 4 had negative coefficients indicating that the more severe their symptoms at a given time, the more likely they are to rate themselves more severely in relation to the clinician QIDS scores. To improve the global model, we would suggest searching for features that might identify this minority of patients who would tend to rate their symptoms lower on average than their peers; such features might be identified using theory and expertise in the study of major depressive disorder.

The Oxford Parkinson's Disease Telemonitoring Dataset comprises a collection of speech signals collected from 42 people (5,875 observations in total) with early-stage Parkinson's Disease (PD) collected during a telemonitoring study to assess progression of PD symptoms remotely using speech characteristics (Tsanas et al., 2010a,b). Previous studies using these dysphonia measures have been able to both distinguish persons with PD from healthy subjects (Little et al., 2009), and to predict PD symptom severity remotely using linear and non-linear regression techniques (Tsanas et al., 2010b). We used gLOP to predict the total Unified Parkinson's Disease Rating Scale (UPDRS) score based on waveform features, using BIC to choose penalization parameters. We included a penalized local intercept for each patient, allowing us to capture variability in average PD symptom severity between patients and to interpret the remaining coefficients as the influence of each feature on the departure from a patient-specific mean severity.

We found that *MDVP absolute jitter*, *MDVP local shimmer (dB)*, *eleven point amplitude perturbation quotient*, *noise-to-harmonics ratio*, *harmonics-to-noise ratio*, *recurrence period density entropy*, *detrended fluctuation analysis*, and *pitch period entropy* were globally predictive; the inclusion of the latter four features is consistent with previous work on classification (Little et al., 2009). The coefficient matrix L was mainly sparse, but all but one feature had one or more local coefficients. Of the 42 patients, 5 appear to be outliers, as the sum of the absolute values of their local coefficients were above the 90th percentile of all values calculated. Based on this, it is possible that including additional features in the model could help to distinguish these patients and future similar patients from the bulk of patients for whom the global model predicts well.

5. Conclusion and Future Work

The gLOP model lays conceptual and methodological groundwork for capturing predictive heterogeneity by identifying predictive outliers, which can help direct future data-gathering activities in cases where data collection may be difficult, invasive, or costly. We may in future want to impose additional constraints on L to allow for only a small number of types of outliers that have the same local coefficients; this could be achieved by shrinking columns of L toward each other, similar in concept to the fused lasso (Tibshirani et al., 2005). Using this approach, we could match new patients to one of a few local "subtypes" of patient in order to achieve better predictions. Another direction for future research is to incorporate post selection inference into gLOP as has been done with LARS (Taylor et al., 2014), which would provide additional confidence information. Finally, we aim to incorporate gLOP into a visual exploratory data analytics system that will reveal predictive outliers and other kinds of hidden structure in datasets used for predictive modelling.

References

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Ali Jalali, Pradeep D Ravikumar, Sujay Sanghavi, and Chao Ruan. A dirty model for multi-task learning. In *NIPS*, volume 3, page 7, 2010.

Max A Little, Patrick E McSharry, Eric J Hunter, Jennifer Spielman, and Lorraine O Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 56(4):1015–1022, 2009.

A. J. Rush, M. Fava, S. R. Wisniewski, P. W. Lavori, M. Trivedi, H. A. Sackeim, and et al. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials*, 25(1):119–42, Feb 2004.

A.John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, Michael E Thase, James H Kocsis, and Martin B Keller. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5):573–583, sep 2003. ISSN 00063223. doi: 10.1016/S0006-3223(02)01866-8. URL `http://www.sciencedirect.com/science/article/pii/S0006322302018668`.

Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. ISSN 1467-9868.

Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Athanasios Tsanas, Max Little, Patrick E McSharry, Lorraine O Ramig, et al. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *Biomedical Engineering, IEEE Transactions on*, 57(4):884–893, 2010a.

Athanasios Tsanas, Max Little, Patrick E McSharry, Lorraine O Ramig, et al. Enhanced classical dysphonia measures and sparse regression for telemonitoring of parkinson's disease progression. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 594–597. IEEE, 2010b.

SJ Wright and J Nocedal. *Numerical optimization*, volume 2. Springer New York, 1999.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

Appendix A

Proof [Of Lemma 1] The proof uses a strategy adapted from Tibshirani et al. (2013) with modifications to accommodate different penalizations on different columns. Let $\hat{\boldsymbol{\xi}}_{\mathcal{A}}$ be the sub-vector of $\hat{\boldsymbol{\xi}}$ containing only the elements with indices in $\mathcal{A}$, and let $\hat{\boldsymbol{\alpha}}_{\mathcal{A}}$ be the analogous sub-vector of $\hat{\boldsymbol{\alpha}}$. Because all coefficients not contained in $\hat{\boldsymbol{\xi}}_{\mathcal{A}}$ must be zero and do not contribute to the lasso fit, equating (8) to the zero vector gives

$$\bar{X}_{\mathcal{A}}^{\mathsf{T}}(\boldsymbol{y} - \bar{X}_{\mathcal{A}}\hat{\boldsymbol{\xi}}_{\mathcal{A}}) = \lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}}. \tag{9}$$

Therefore $\lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}} \in \text{row}(\bar{X}_{\mathcal{A}})$, and $\bar{X}_{\mathcal{A}}^{\mathsf{T}}(\bar{X}_{\mathcal{A}}^{\mathsf{T}})^{+}\lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}} = \lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}}$ where X^{+} indicates the Moore-Penrose pseudoinverse of X. Further algebraic manipulation of (9) gives

$$\bar{X}_{\mathcal{A}}\hat{\boldsymbol{\xi}}_{\mathcal{A}} = \bar{X}_{\mathcal{A}}\left[\bar{X}_{\mathcal{A}}^{+}(\boldsymbol{y} - (\bar{X}_{\mathcal{A}}^{\mathsf{T}})^{+}\lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}})\right].$$

Note that $\bar{X}_{\mathcal{A}}\hat{\boldsymbol{\xi}}_{\mathcal{A}}$ is the optimal lasso **fit**, which is unique even though the coefficients providing that fit may not be (Tibshirani et al., 2013). The set of all optimal coefficient vectors is given by

$$\{\boldsymbol{\xi}_{\mathcal{A}} : \boldsymbol{\xi}_{\mathcal{A}} = \bar{X}_{\mathcal{A}}^{+}(\boldsymbol{y} - (\bar{X}_{\mathcal{A}}^{\mathsf{T}})^{+}\lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}}) + \mathbf{z}\}$$

for $\mathbf{z} \in \text{null}(\bar{X}_{\mathcal{A}})$ and $\text{sgn}(\bar{X}_{\mathcal{A}}^{+}(\boldsymbol{y} - (\bar{X}_{\mathcal{A}}^{\mathsf{T}})^{+}\lambda\hat{\boldsymbol{\alpha}}_{\mathcal{A}}) + \mathbf{z}) = \hat{\boldsymbol{\alpha}}_{\mathcal{A}}$. If $\bar{X}_{\mathcal{A}}$ has full rank then $\text{null}(\bar{X}_{\mathcal{A}}) = \mathbf{0}$ and we have a unique optimal coefficient vector obtained by setting $\mathbf{z} = \mathbf{0}$. ■

Proof [Of Lemma 3] Suppose that $\bar{X}_{\mathcal{A}}$ does not have full rank. Then for some column i of $\bar{X}_{\mathcal{A}}$, there exist weights c_j such that

$$\bar{X}_{\mathcal{A},i} = \sum_{j \neq i} c_j \bar{X}_{\mathcal{A},j}. \tag{10}$$

Recall that each index has a corresponding $\alpha_{\mathcal{A},i} \in \{1, -1\}$ as defined in (8). It follows that

$$\bar{X}_{\mathcal{A},i} = \sum_{j \neq i} (c_j \alpha_{\mathcal{A},i}\alpha_{\mathcal{A},j}) \cdot \frac{\alpha_{\mathcal{A},j}}{\alpha_{\mathcal{A},i}} \bar{X}_{\mathcal{A},j}.$$

By definition, $\frac{\alpha_{\mathcal{A},j}}{\alpha_{\mathcal{A},i}} \in \{-1, 1\}$. We will now show that the weights $c_j\alpha_{\mathcal{A},i}\alpha_{\mathcal{A},j}$ sum to 1. Recall from (9) that $(\bar{X}_{\mathcal{A},i})^{\mathsf{T}}(\boldsymbol{y} - \bar{X}\hat{\boldsymbol{\xi}}) = \lambda_i\alpha_{\mathcal{A},i}$. Therefore, using (10) we have

$$\begin{aligned}
(\bar{X}_{\mathcal{A},i})^{\mathsf{T}}(\boldsymbol{y} - \bar{X}\hat{\boldsymbol{\xi}}) &= \sum_{j \neq i} c_j (\bar{X}_{\mathcal{A},i}^{\mathsf{T}})(\boldsymbol{y} - \bar{X}\hat{\boldsymbol{\xi}}) \\
\alpha_{\mathcal{A},i} &= \sum_{j \neq i} c_j \alpha_{\mathcal{A},j} \\
1 &= \sum_{j \neq i} c_j \alpha_{\mathcal{A},i}\alpha_{\mathcal{A},j}
\end{aligned}$$

This establishes that if $\bar{X}_{\mathcal{A}}$ does not have full rank, then its columns are not AIN because we can produce signs $s_i = \frac{\alpha_{\mathcal{A},j}}{\alpha_{\mathcal{A},i}}$ and weights $w_i = c_j\alpha_{\mathcal{A},i}\alpha_{\mathcal{A},j}$ as witnesses. ■

Appendix B

In order to evaluate gLOP in a controlled fashion and to compare it with the dirty model, we conduct experiments with varying numbers of features and patients. We draw elements of the design matrices from a Gaussian with 0 mean and unit variance, and we generate observations $\boldsymbol{y}^k$ by adding Gaussian noise (0 mean, unit variance) to $X\boldsymbol{\theta}^k$ for each of κ patients, where $\boldsymbol{\theta}^k$ gives the true model coefficients for that patient.

First, we explored differences in the output of gLOP versus the dirty model using a small example ($p = 4, \kappa = 5, n = 64$) with the following parameters:

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^2 = \boldsymbol{\theta}^3 = (0, 0, 3, 3)^\mathsf{T}, \quad \boldsymbol{\theta}^4 = (0, 0, -3, 3)^\mathsf{T}, \quad \boldsymbol{\theta}^5 = (0, 3, 0, 3)^\mathsf{T}$$

Identical data sets were used for each algorithm with $\lambda_{\boldsymbol{g}/B} = 5$ and $\lambda_{L/S} = 10$ was chosen by cross-validation. Based on this example, we observed that gLOP's $\hat{\boldsymbol{g}} + \hat{L}_k$ recovers the true model parameters for patient k quite closely, although false inclusions were present in two of the patients. In contrast, the parameters recovered by the dirty model did not capture any variation between patients in $\hat{S}$, which was exactly zero, and induced little variation between patients in B. We found that in our experiments it was often impossible to find a pair of λ_B and λ_S such that both B and S contained non-zero values.

We also conduct larger-scale experiments to evaluate predictive performance; in these we use three different patient-types with true parameters $\boldsymbol{\tau}^1$, $\boldsymbol{\tau}^2$, and $\boldsymbol{\tau}^3$, given by

$$\boldsymbol{\tau}^1 = (\underbrace{3, ..., 3}_{1,...,\frac{p}{4}}, \underbrace{0, ..., 0}_{\frac{p}{4}+1,...,p})^\mathsf{T}, \boldsymbol{\tau}^2 = (\underbrace{3, -3, ..., 3, -3}_{1,...,\frac{p}{2}}, \underbrace{0, ..., 0}_{\frac{p}{2}+1,...,p})^\mathsf{T}, \boldsymbol{\tau}^3 = (\underbrace{-3, 3, ..., -3, 3}_{1,...,\frac{p}{8}}, \underbrace{0, ..., 0}_{\frac{p}{8}+1,...,p})^\mathsf{T}.$$

Note that 3/4 of the entries in $\boldsymbol{\tau}^1$ are zero, $\boldsymbol{\tau}^2$ is less sparse than $\boldsymbol{\tau}^1$, and $\boldsymbol{\tau}^3$ is more sparse than $\boldsymbol{\tau}^1$. In each experiment, $\frac{\kappa}{8}$ patients are generated using each of $\boldsymbol{\tau}_2$ and $\boldsymbol{\tau}_3$; the remaining $\kappa - \frac{\kappa}{4}$ patients are generated using $\boldsymbol{\tau}_1$.

To choose values for λ_g and λ_L, we perform 10-fold cross validation over a grid of points $(\lambda_{\boldsymbol{g}}, \lambda_L)$ but constrain the results to include only cases where $\lambda_{\boldsymbol{g}} \leq \lambda_L$. The folds are stratified across patients. The grid ranges from 0 to 100 in steps of 5 along each dimension. While this may appear coarse, note that our loss function (squared error) is not normalized by n as is sometimes common, in order to compare more directly with the original dirty model implementation. Thus the range of useful λ is wider in our experiments than in other lasso applications. Once CV error for all pairs has been calculated, the pair of $\lambda_{\boldsymbol{g}}$ and λ_L with the lowest error is selected. We break ties in favour of larger λ_L and then larger $\lambda_{\boldsymbol{g}}$ in order to obtain the sparsest model. We evaluate the prediction error for each learned model using a large ($n = 1000$) held-out test set.

Appendix C

Table 1: Test error results for gLOP versus the dirty model. All errors attained by the dirty model and the lasso were statistically significantly worse than those of gLOP ($p < 0.05$ on an independent two-sample t-test).

p	κ	n	gLOP	Dirty M.	Lasso
16	16	64	1.3931	6.3718	39.5652
			$\pm$0.0637	$\pm$0.2599	$\pm$0.5335
16	128	64	1.4602	2.2171	39.8316
			$\pm$0.0237	$\pm$0.121	$\pm$0.1668
128	16	64	93.6959	141.1617	306.5222
			$\pm$7.5097	$\pm$9.5854	$\pm$ 3.6947
128	128	64	73.9881	141.1624	307.3203
			$\pm$2.7155	$\pm$3.7506	$\pm$1.2835