Optimal Constraint-Based Regularization for Parameter Estimation Problems

Caroline Satye Martins Nakama‡, Galo Antonio Carrillo Le Roux‡, and Victor M. Zavala¶∗

‡Department of Chemical Engineering
Polytechnic School, University of São Paulo, São Paulo, Brazil
¶Department of Chemical and Biological Engineering
University of Wisconsin-Madison, Madison, Wisconsin, USA

Abstract

We explore strategies to regularize ill-posed parameter estimation problems by using constraints. Constraints provide a flexible and easily interpretable approach to reduce the parameter space (compared to objective-based regularization). We begin our discussion by revisiting strategies to exploit the eigenvalue structure of the Hessian matrix (inverse of the parameter covariance matrix) to design optimal constraints that minimize the variance of the parameter estimates. We show that these strategies can be used in more general settings by exploiting information from the reduced Hessian matrix. We also derive a elastic net strategy to sparsify the constraints and facilitate their interpretability. Our analysis also highlights that subspace parameter selection can be interpreted as a constraint regularization technique that is easy to interpret but that is suboptimal.

Keywords: parameter estimation; ill-conditioning; regularization; eigenvalues; constraints; elastic net

1 Introduction

Parameter estimation problems are often ill-conditioned due to the lack of sufficient experimental data, due to model overparameterization, or due to large measurement errors. Ill-conditioning manifests as high sensitivity of the parameter estimates to data and as high parameter variance. Examples of applications where ill-conditioning arises include image restoration [1], gene selection [2], gas emission source identification [3], and seasonal forecasting [4]. Ill-conditioning can be addressed by using regularization straitgies. Such strategies can be implemented by using objective penalization terms or constraints. Tikhonov and ridge regression, for instance, use a penalization term in the form of an $\ell_2$ norm to stabilize the parameters [5]. Lasso regularization uses an $\ell_1$ norm and the elastic net uses a combination $\ell_1$ and $\ell_2$ norms [6, 7]. An issue associated with objective penalization is that tuning the regularization parameter is usually non-trivial [8].

*Corresponding Author: victor.zavala@wisc.edu
Ill-conditioning can also be addressed by enforcing constraints on the parameters. Constraints have the effect of reducing the allowable parameter space to be explored. A straightforward approach to reduce the parameter space is simply to fix a subset of parameters (a strategy called parameter subset selection) but finding an optimal set of parameters is difficult (this is a combinatorial problem). Another popular strategy consists of using trust-region constraints, which allows setting a region (a subspace) over which parameters can be searched for [9]. This approach, however, induces similar behavior to objective regularization [10, 11]. For instance, adding an $\ell_2$ norm constraint [12] induces similar behavior to Tikhonov and ridge regression [5]. Park [13] showed that one can construct optimal constraints that minimize the parameter covariance by using eigenvectors of the kernel (Hessian) matrix. Principal component regression [14] is another powerful approach that reduces the parameter subspace by using eigenvectors of the Hessian matrix.

In this work, we revisit classical results to derive a unifying framework for constrained-based regularization for linear estimation problems. The framework reveals a strategy to implement optimal constraints in a general setting by using eigenvector information of the reduced Hessian. This also reveals that eigenvectors provide optimal linear combinations of parameters (clusters) to be estimated. In other words, this regularization approach finds optimal clusters of parameters. However, we note that these clusters are dense and difficult to interpret. Motivated by this limitation, we propose an elastic net formulation that sparsifies the eigenvectors in order to construct sparse clusters that are close to optimal and more interpretable. This sparsification approach also reveals that parameter subset selection can be interpreted as a limiting case of using eigenvector constraints; consequently, subset selection is inherently suboptimal. Our analysis also implies that standard trust-region constraints are inherently suboptimal. We provide examples that illustrate the concepts and discuss how to use the gained insights to address nonlinear estimation problems.

This paper is organized as follows. Section 2 states the estimation problem under study. Section 3 discusses strategies and benefits of constrained-based regularization strategies and discuss extensions. Lastly, there is a brief discussion on applicability to nonlinear problems. In section 5 we illustrate the properties of eigenvalue based regularization with two synthetic case studies, one with near collinearities in the input data and one with rank deficiency. A third case study is presented combining regularization methods with an approach to estimate nonlinear models using linear regression when insufficient data is available.

2 Estimation Problem

We introduce basic terminology and notation by considering an unconstrained linear estimation problem. The goal is to estimate a set of parameters $\theta \in \mathbb{R}^m$ for a linear response model $\eta_k = \theta^T x_k + \epsilon_k$ from a set of input observations $x_k \in \mathbb{R}^m$ and output observations $\eta_k \in \mathbb{R}$, where the observation index is given by $k = 1, ..., n$. Here, $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. random variables that capture unknown behavior. The response for all the observations is expressed in the compact form as $\eta = X \theta + e$, where $X = (x_1, ..., x_n) \in \mathbb{R}^{n \times m}$ is the input data matrix, $\eta = (\eta_1, ..., \eta_m) \in \mathbb{R}^m$ is the response data vector, and $e \sim \mathcal{N}(0, \sigma^2 I)$.
The model parameters are estimated by solving the quadratic programming (QP) problem:

$$\hat{\theta} \in \arg \min_{\theta} \frac{1}{2} (\eta - X\theta)^T (\eta - X\theta).$$  \hspace{1cm} (2.1)

The first-order conditions of this QP lead to the solution:

$$\hat{\theta} = K^{-1} X^T \eta.$$  \hspace{1cm} (2.2)

Here, \(K := X^T X \in \mathbb{R}^{m \times m}\) is the kernel matrix, which is symmetric and encodes all information of the input data and associated impact on the estimated parameters. Our setting also allows for nonlinear estimation settings in which the data matrix \(X\) is transformed using basis functions [15].

Given that \(\eta = X\theta + e\) and \(e \sim \mathcal{N}(0, \sigma^2 I)\), the estimated parameters \(\hat{\theta}\) are a linear transformation of a Gaussian variable with covariance matrix:

$$\mathbb{V}[\hat{\theta}] = \sigma^2 K^{-1}$$  \hspace{1cm} (2.3)

From (2.2) it is clear that the estimated parameters exist, are unique, and provide a minimum of the QP if and only if the input data matrix \(X\) has full rank (in which case the kernel matrix \(K\) is nonsingular and positive definite). From (2.3) it is clear that the covariance matrix of the parameters is linked to the inverse of the kernel matrix \(K\). To make this connection clearer, we express the kernel matrix using an eigenvalue decomposition of the form \(K = V \Lambda V^T\) where \(V = (v_1, \ldots, v_m)\) is the eigenvector matrix with columns \(v_k \in \mathbb{R}^m\) and \(\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)\) is the eigenvalue matrix with entries \(\lambda_k \in \mathbb{R}\).

The eigenvalue decomposition allows us to express the kernel matrix as \(K = \sum_{k=1}^{m} \lambda_k v_k v_k^T\) and to compute its Frobenius norm as \(\|K\|_F = \|\Lambda\|_F = \sqrt{\sum_{k=1}^{m} \lambda_k^2}\). Moreover, because the eigenvector matrix is orthogonal, we have that \(V^{-1} = V^T\) and thus \(K^{-1} = (V^T \Lambda V)^{-1} = V^T \Lambda^{-1} V = \sum_{k=1}^{m} \lambda_k^{-1} v_k v_k^T\).

We thus have that small eigenvalues of the kernel matrix \(K\) give rise to large eigenvalues of the covariance matrix \(\mathbb{V}[\hat{\theta}]\) and thus to a large Frobenius norm \(\|\mathbb{V}[\hat{\theta}]\|_F = \sigma^2 \sqrt{\sum_{k=1}^{m} \lambda_k^{-2}}\). Because the eigenvalues of \(\mathbb{V}[\hat{\theta}]\) provide a measure of the size of the confidence levels of the parameters, we have that confidence decreases as the eigenvalues of \(K\) decrease.

An ill-conditioned estimation problem is one in which the estimated parameters exhibit high variance (variability) because the available data is insufficient to resolve them with high confidence. In the extreme degenerate case in which \(X\) does not have full rank we have that \(K\) has zero eigenvalues (is singular) and the estimated parameters are not unique and have infinite variance. In such a case, we can use a singular value decomposition (SVD) of \(X\) to obtain eigenvalue and eigenvector information for \(K\). The SVD decomposition is \(X = USV^T\) where \(S = \text{diag}(s_1, \ldots, s_m)\) are the singular values of \(X\) and \(U, V\) are orthogonal matrices. We thus have that:

$$K = X^T X$$
$$= (USV^T)^T (USV^T)$$
$$= VS^2 V^T$$  \hspace{1cm} (2.4)

Consequently, we have that \(\Lambda = S^2\) are the eigenvalues of \(K\) and \(V\) are its eigenvectors.
When data is insufficient to resolve the parameters with sufficient confidence (or at all) it is necessary to devise a strategy to regularize the problem. A popular strategy to achieve this consists of using *objective regularization* terms. For instance, the elastic net formulation has become a standard regularization approach [7]:

\[
\hat{\theta} \in \arg \min_{\theta} \frac{1}{2}(\eta - X\theta)^T(\eta - X\theta) + \kappa_1\|\theta\|_1 + \kappa_2\|\theta\|_2^2.
\]

(2.5)

Here, \(\kappa_1, \kappa_2 \in \mathbb{R}_+\) are tuning weights for the \(\ell_1\) norm \(\|\cdot\|_1\) and the squared \(\ell_2\) norm \(\|\cdot\|_2^2\). It can be easily shown that the squared \(\ell_2\) norm has the effect of displacing the eigenvalues of the kernel matrix \(K = (X^TX + \kappa_2I)\). In particular, one can use an eigenvalue decomposition to show that the inverse kernel matrix is

\[
K^{-1} = \sum_{k=1}^m (\lambda_k + \kappa_2)^{-1}v_kv_k^T
\]

and with this prevent damaging effects of small eigenvalues of \(X^TX\). The \(\ell_1\) norm has the effect of reducing the number of nonzero parameters (it sparsifies the estimate. As we will see, an issue with objective regularization is that it provides limited control over parameter behavior.

### 3 Constraint-Based Regularization Strategies

In this work, we seek to control parameter behavior by using *constraints*. We begin the discussion by considering the constrained estimation problem:

\[
\hat{\theta} \in \arg \min_{\theta} \frac{1}{2}(\eta - X\theta)^T(\eta - X\theta) \quad \text{ s.t. } R\theta = r,
\]

(3.6a)

(3.6b)

Here \(R \in \mathbb{R}_{p \times m}\) is a given constraint matrix and \(r \in \mathbb{R}^p\) is a given constraint right-hand side vector. Clearly, the estimation problem is an equality-constrained QP with \(m - p\) degrees of freedom. Consequently, inducing constraints has the effect of reducing the parameter subspace (from \(m\) to \(m - p\)).

The constraint matrix and right-hand side vector can be used to control behavior of the estimated parameter. For instance, consider a simple example with \(m = 3\) parameters \(\theta = (\theta_1, \theta_2, \theta_3)\). Assume that we would like to fix a subset of parameters (say \(\theta_1\) and \(\theta_3\)) to given values \(r_1\) and \(r_2\), respectively. This can be achieved by imposing the constraints:

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\theta_1 \\
\theta_2 \\
\theta_3
\end{bmatrix} =
\begin{bmatrix}
 r_1 \\
r_2
\end{bmatrix}.
\]

(3.7)

In this case, we have \(p = 2\) and thus the problem has \(m - p = 1\) degrees of freedom left (corresponding to \(\theta_1\)). Fixing the parameters to specific values is known as *parameter subset selection* and has the goal of reducing parameter variance by reducing the parameter space. An issue with this approach is that selecting a suitable set of parameters to fix is non-trivial (a combinatorial number of parameter sets can be fixed). In particular, imagine that we would like to identify the set of \(p\) parameters to fix in order to minimize variance \(\nabla \hat{\theta}\). Subset selection can thus be interpreted as finding the non-zero locations for the constraint coefficient matrix \(R\) that achieves minimum variance (with the implicit restriction of having only one non-zero entry per row).
Consider again the $m = 3$ parameter problem but assume now that the parameters are related as $R_{11}\theta_1 + R_{12}\theta_2 + 2 + R_{13}\theta_3 = r_1$ and $R_{21}\theta_1 + R_{22}\theta_2 + R_{23}\theta_3 = r_2$. This can be modeled using the constraints:

\[
\begin{bmatrix}
R_{11} & R_{12} & R_{13} \\
R_{21} & R_{22} & R_{23}
\end{bmatrix}
\begin{bmatrix}
\theta_1 \\
\theta_2 \\
\theta_3
\end{bmatrix}
=
\begin{bmatrix}
r_1 \\
r_2
\end{bmatrix}.
\tag{3.8}
\]

In this case, we have $p = 2$ and thus the problem has also $m - p = 1$ degrees of freedom but, compared to subset selection (the degree of freedom can be either $\theta_1$, $\theta_2$, or $\theta_3$). We thus have that the constraint set provides more flexibility than fixing any $p = 2$ combination of $\theta_1$, $\theta_2$, or $\theta_3$ using subset selection. In fact, note that the values of the estimated parameters are not relevant as long as they satisfy the constraints. As with subset selection, imagine now that we would like to find the constraint matrix coefficients $R_{ij}$ that achieve minimum parameter variance. This approach is more flexible than subset selection because it does not pre-specifies a sparsity pattern on the constraint matrix. We will see that finding the optimal matrix $R$ can be found in a tractable manner.

To gain some intuition into how to design constraints that minimize variance of parameter estimates, we consider the first-order optimality conditions of the constrained QP (3.6):

\[
X^T X\theta + R^T \lambda - X^T \eta = 0
\tag{3.9a}
\]

\[
R\theta - r = 0
\tag{3.9b}
\]

where $\lambda \in \mathbb{R}^p$ are Lagrange multipliers. Multiplying the first condition through by $K^{-1}$ we obtain:

\[
\theta = K^{-1} X^T \eta - K^{-1} R^T \lambda
\tag{3.10}
\]

Inserting into the second condition we obtain:

\[
RK^{-1} X^T \eta - RK^{-1} R^T \lambda - r = 0
\tag{3.11}
\]

and thus,

\[
\lambda = (RK^{-1} R^T)^{-1} RK^{-1} X^T \eta - (RK^{-1} R^T)^{-1} r.
\tag{3.12}
\]

We thus have that:

\[
\hat{\theta} = K^{-1} X^T \eta - K^{-1} R^T (RK^{-1} R^T)^{-1} RK^{-1} X^T \eta + K^{-1} R^T (RK^{-1} R^T)^{-1} r.
\tag{3.13}
\]

Now recall that the unconstrained parameter estimate is given by $\hat{\theta} = K^{-1} X^T \eta$ and thus:

\[
\tilde{\theta} = \hat{\theta} - K^{-1} R^T (RK^{-1} R^T)^{-1} R\hat{\theta} + K^{-1} R^T (RK^{-1} R^T)^{-1} r = \Gamma \hat{\theta} + \tilde{r},
\tag{3.14}
\]

where $\Gamma := I - K^{-1} R^T (RK^{-1} R^T)^{-1} R$ and $\tilde{r} := K^{-1} R^T (RK^{-1} R^T)^{-1} r$. Consequently, we have that:

\[
\nabla \tilde{\theta} = \nabla \Gamma \hat{\theta} + \nabla \tilde{r}
= \Gamma \nabla \hat{\theta} \Gamma^T
= \sigma^2 K^{-1} - \sigma^2 K^{-1} R^T (RK^{-1} R^T)^{-1} RK^{-1}.
\tag{3.15}
\]
From these derivations we observe that the constraint right-hand side \( r \) influences the value of the estimated parameters but does not affect the parameter covariance. Moreover, the covariance \( \nabla[\hat{\theta}] \) can indeed be controlled by selecting a suitable constraint matrix \( R \). We now proceed to derive strategies to select an optimal constraint matrix.

### 3.1 Regularization using Eigenvector Constraints

Park [13] noticed that a matrix \( R \) (of rank \( p \leq m \)) that minimizes covariance can be obtained from the eigenvalue decomposition of the kernel matrix \( K \). To see how this is done, we write the eigenvalue decomposition of the kernel matrix as:

\[
K = [V_1 | V_2] \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2 \\ \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \\ \end{bmatrix}
\] (3.16)

where \( \Lambda_1 = (\lambda_1, \lambda_2, ..., \lambda_p) \) contains the \( p \)-largest eigenvalues of \( K \) with associated eigenvectors \( V_1 = (v_1, v_2, ..., v_p) \in \mathbb{R}^{m \times p} \) and \( \Lambda_2 = (\lambda_{p+1}, \lambda_{p+2}, ..., \lambda_m) \) contains the \( m - p + 1 \) smallest eigenvalues with eigenvectors \( V_2 = (v_{t+1}, v_{t+2}, ..., v_m) \in \mathbb{R}^{m \times (m-p+1)} \). We note that \( K^{-1} = V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T \) and we recall that \( V_1 \) and \( V_2 \) are orthogonal and thus \( V_1^TV_2 = 0 \) and \( V_2^TV_1 = 0 \) and \( V_2^TV_2 = I \). Assume now that \( R = \sqrt{\Lambda_2}V_2^T \) and, upon substitution in (3.15), we observe that:

\[
R K^{-1} R^T = \sqrt{\Lambda_2}V_2^T (V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T)V_2\sqrt{\Lambda_2}
= \sqrt{\Lambda_2}V_2^T V_2\Lambda_2^{-1}V_2^T V_2\sqrt{\Lambda_2}
= \begin{bmatrix} \Lambda_2 \end{bmatrix}
\]

and thus:

\[
K^{-1} R^T (R K^{-1} R^T)^{-1} R K = (V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T)V_2\sqrt{\Lambda_2}\sqrt{\Lambda_2}V_2^T (V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T)
= (V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T)V_2\Lambda_2^{-1}V_2^T (V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T)
= V_2V_2^T (V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T)
= V_2\Lambda_2^{-1}V_2^T.
\]

Merging terms we obtain:

\[
\nabla[\hat{\theta}] = \sigma^2(V_1\Lambda_1^{-1}V_1^T + V_2\Lambda_2^{-1}V_2^T) - \sigma^2V_2\Lambda_2^{-1}V_2^T
= \sigma^2(V_1\Lambda_1^{-1}V_1^T).
\] (3.17)

Consequently, we see that enforcing the constraint \( R\theta = r \) (with \( R = \sqrt{\Lambda_2}V_2^T \)) minimizes the effect of the damaging (small) eigenvalues of \( K \) on the covariance matrix; in particular, we note that \( \|\nabla[\hat{\theta}]\|_F = \sigma^2 \sum_{k=1}^{p} \lambda_k^{-2} \). In other words, the constraint matrix \( R \) is optimal.

### 3.2 Regularization with Principal Component Regression

We NOW show that using eigenvector constraints induces the same regularization effect as principal component (PC) regression [14]. This reveals interesting insights on the effect of the optimal eigenvalue
constraints. In PC regression, the data matrix $X$ is projected onto the space of the $p$ largest eigenvectors as $XV_1 = (Xv_1, \ldots, Xv_p) \in \mathbb{R}^{n \times p}$ and the estimation is performed using the reduced data matrix. In other words, the data is reduced to its leading $p$ principal components. PC regression is equivalent to define a reduced parameter vector (of dimension $p \leq m$) of the form $\gamma = V_1^T \theta \in \mathbb{R}^p$ and to solve the estimation problem over such reduced space:

$$\hat{\gamma} \in \arg \min_{\gamma} \frac{1}{2}(\eta - XV_1 \gamma)^T(\eta - XV_1 \gamma). \quad (3.18)$$

A solution in the original parameter space is recovered from $\hat{\theta} = V_1 \hat{\gamma}$. Clearly, the above QP is equivalent to the constrained QP:

$$\begin{align*}
(\hat{\theta}, \hat{\gamma}) & \in \arg \min_{\gamma, \theta} \frac{1}{2}(\eta - X\theta)^T(\eta - X\theta) \quad (3.19a) \\
& \text{s.t. } V_1^T \theta = \gamma. \quad (3.19b)
\end{align*}$$

We thus note that PC regression can be expressed as a constraint-based regularization formulation with coefficient matrix $R = V_1^T$ but with variable right-hand side $r = \gamma$ (the right-hand side of the constraint does not need to be specified). Now note that the reduced parameter solution is given by $\hat{\gamma} = (V_1^T X^T X V_1)^{-1} V_1^T X^T \eta$. This highlights that PC regression is equivalent to perform estimation on the reduced data set $V_1 X$. Moreover, the covariance of $\hat{\gamma}$ is given by:

$$\begin{align*}
\mathbb{V}[\hat{\gamma}] &= \sigma^2 (V_1^T X^T X V_1)^{-1} \\
&= \sigma^2 (V_1^T (V_1 \kappa_1 V_1^T + V_2 \kappa_2 V_2^T) V_1)^{-1} \\
&= \sigma^2 \Lambda_1^{-1}, \quad (3.20)
\end{align*}$$

where we exploit the orthogonality of the eigenvectors (i.e., $V_2^T V_1 = 0$ and $V_1^T V_1 = I$). This indicates that the reduced parameters are uncorrelated. The covariance of the full parameter estimate $\hat{\theta}$ is:

$$\begin{align*}
\mathbb{V}[\hat{\theta}] &= V_1 \mathbb{V}[\hat{\gamma}] V_1^T \\
&= \sigma^2 V_1 \Lambda_1^{-1} V_1^T. \quad (3.21)
\end{align*}$$

This reveals that eigenvector constraints and PC regression induce the same regularization effect (they eliminate the effect of the small eigenvalues from the kernel matrix) but they follow different implementation mechanisms. Specifically, PC regression creates a reduced set of parameters $\gamma = V_1^T \theta$ that are linear combinations of the original parameters (these can be interpreted as parameter clusters) and seeks to find cluster values $\hat{\gamma}$ that minimize the model error. Eigenvectors of the kernel matrix provide the coefficients for the clusters that minimize the parameter covariance. This scheme thus provides a mechanism to optimally cluster parameters when individual parameters cannot be estimated with high confidence given the available data. This clustering effect is less obvious when using eigenvector constraints.

An important property of PC regression is that only eigenvector information is needed to perform regularization (the eigenvalues are not needed, as in the regularization approach with eigenvector constraints). This property will become relevant when seeking for sparse approximations of eigenvectors that can help interpret parameter clusters.
3.3 Regularization with Sparse Principal Component Regression

Our previous discussion reveals that eigenvector constraints are optimal and this regularization approach is equivalent to PC regression. We have also seen that PC regression can be interpreted as estimating an optimal set of parameter clusters \( \gamma = V_1^T \theta \). Unfortunately, the coefficients of these clusters (the eigenvectors \( V_1 \)) are dense; because of this, every cluster depends on all parameters (i.e., \( \gamma_i = \sum_{j=1}^{m} v_{ij} \theta_j \) for \( i = 1, \ldots, p \)). As a result, the clusters might be difficult to interpret. We are thus interested in computing sparse versions of the \( p \) leading eigenvectors \( V_1 = (v_1, \ldots, v_p) \), which we denote as \( \tilde{V}_1 = (\tilde{v}_1, \ldots, \tilde{v}_p) \) to obtain sparse clusters \( \gamma = \tilde{V}_1^\theta \). This can be achieved by using an elastic net approach (also known as sparse PC). In [16], it is shown that sparse approximations of the leading eigenvectors of \( K = X^T X \) can be obtained from the solution of the elastic net problem:

\[
\tilde{v}_j \in \arg \min_v \frac{1}{2} \|Xv_j - Xv\|_2^2 + \kappa_2 \|v\|_2^2 + \kappa_1 \|v\|_1 \tag{3.22}
\]

for \( j = 1, \ldots, p \) and \( \tilde{v}_j \leftarrow \tilde{v}_j / \|\tilde{v}_j\|_2 \). This approach is derived based on the observation that the \( p \) leading eigenvectors \( v_j, j = 1, \ldots, p \) of \( X^T X \) can be recovered from the solution of the problem:

\[
\tilde{v}_j = \arg \min_v \frac{1}{2} \|Xv_j - Xv\|_2^2 + \kappa_2 \|v\|_2^2 \tag{3.23}
\]

for any value of \( \kappa_2 \in \mathbb{R}_+ \). Here, the tuning parameter \( \kappa_1 \) is used to control the sparsity of the eigenvectors \( \tilde{V}_1 \). The idea of this approach (in our estimation context) is to obtain sparse clusters \( \gamma = \tilde{V}_1^\theta \) that more clearly reveal dominant parameters.

We note that the sparse PC approach does not deliver eigenvalue information (it only adjusts the eigenvectors). This, however, does not represent an issue because we have seen that one can implement optimal constraint regularization in the form of PC regression. A larger issue with using the sparse PC approach, however, is that the sparse eigenvectors might fail to reduce the parameter variance (which is our main goal). To see this, we note that the PC regression problem with sparse eigenvectors is given by:

\[
\tilde{\gamma} \in \arg \min_{\gamma} \frac{1}{2} (\eta - X\tilde{V}_1 \gamma)^T (\eta - X\tilde{V}_1 \gamma), \tag{3.24}
\]

and with \( \tilde{\theta} = \tilde{V}_1 \tilde{\gamma} \). Now note that the parameter covariance is given by:

\[
\nabla[\tilde{\theta}] = \sigma^2 \tilde{V}_1 (\tilde{V}_1^T X^T X \tilde{V}_1)^{-1} \tilde{V}_1^T \\
= \sigma^2 \tilde{V}_1 (\tilde{V}_1^T (V_1 \kappa_1 V_1^T + V_2 \kappa_2 V_2^T) \tilde{V}_1)^{-1} \tilde{V}_1^T. \tag{3.25}
\]

We note that \( \tilde{V}_1 \) is not orthogonal to \( V_2 \) and thus the effect of the small (damaging) eigenvalues of \( K \) is not eliminated. Moreover, we want \( \tilde{V}_1^T V_1 \approx I \). Motivated by this observation, we propose the following elastic net formulation with orthogonality constraints:

\[
\tilde{v}_j = \arg \min_v \frac{1}{2} \|Xv_j - Xv\|_2^2 + \kappa_2 \|v\|_2^2 + \kappa_1 \|v\|_1 \tag{3.26a}
\]

\[
\text{s.t. } v^Tv_i = 0, \ i = p + 1, \ldots, m \quad \text{and} \quad i \neq j \tag{3.26b}
\]

\[
v^Tv_i = 1, \ i = j \tag{3.26c}
\]
which is solved for \( j = 1, \ldots, p \) and with \( \tilde{v}_j \leftarrow \tilde{v}_j/\|\tilde{v}_j\|_2 \) to construct the sparse eigenvector matrix \( \tilde{V}_1 = (\tilde{v}_1, \ldots, \tilde{v}_p) \). We will see that this approach delivers sparse clusters while achieving close-to-optimal reductions in parameter variance.

4 Regularization for General Formulations

We now show how to design optimal regularization constraints in a general estimation setting. Consider the estimation problem:

\[
(\hat{w}, \hat{\theta}) \in \arg \min_{w, \theta} \frac{1}{2} \sum_{k=1}^{n} (\eta_k - C_k w_k)^T (\eta_k - C_k w_k) \tag{4.27a}
\]

\[
\text{s.t. } A_k w_k + B_k \theta = 0, \ k = 1, ..., n. \tag{4.27b}
\]

Here, \( \theta \in \mathbb{R}^m \) are the parameters, \( w_k \in \mathbb{R}^q \) are hidden variables (e.g., state variables) for observation \( k \), and matrices \( A_k \in \mathbb{R}^{q \times q} \) and \( B_k \in \mathbb{R}^{q \times m} \) capture relationships (physical or empirical) between the hidden variables and the parameters in observation \( k \). We do not specify a right-hand side for this constraint in order to simplify the analysis (we have seen that the right-hand side is irrelevant). The matrix \( C_k \in \mathbb{R}^{t \times q} \) maps the state to the response variable \( \eta_k \in \mathbb{R}^t \). We also define the state vector \( w := (w_1, ..., w_n) \in \mathbb{R}^{n \cdot q} \) and response vector \( \eta = (\eta_1, ..., \eta_n) \in \mathbb{R}^{n \cdot t} \).

The above estimation problem can be posed in the following compact form:

\[
(\hat{w}, \hat{\theta}) \in \arg \min_{w, \theta} \frac{1}{2} (\eta - C w)^T (\eta - C w) \tag{4.28a}
\]

\[
\text{s.t. } A w + B \theta = 0. \tag{4.28b}
\]

Here, matrices \( A \in \mathbb{R}^{n \cdot q \times n \cdot q} \), \( B \in \mathbb{R}^{n \cdot q \times m} \), and \( C \in \mathbb{R}^{n \cdot t \times n \cdot q} \) are constructed from the submatrices \( A_k, B_k, \) and \( C_k \), respectively. Typically, \( n \cdot q \geq m \) because the state space is typically larger than the parameter space and variable vector \( w \) embeds state variables for all observations.

The first-order conditions of the estimation problem are given by:

\[
C^T C w + A^T \nu - C^T \eta = 0 \tag{4.29a}
\]

\[
B^T \nu = 0 \tag{4.29b}
\]

\[
A w + B \theta = 0. \tag{4.29c}
\]

This can be expressed in matrix form as:

\[
\begin{bmatrix}
C^T C & A^T \\
B^T & A & B
\end{bmatrix}
\begin{bmatrix}
w \\
\theta \\
\nu
\end{bmatrix}
= 
\begin{bmatrix}
C^T \eta \\
0 \\
0
\end{bmatrix} \tag{4.30}
\]

Assuming that \( C^T C \) is invertible (e.g., \( C \) has full rank), we can perform block elimination to obtain:

\[
w = (C^T C)^{-1}(-A^T \nu + C^T \eta) \tag{4.31a}
\]

\[
\nu = (A(C^T C)^{-1}A^T)^{-1}(B \theta + A(C^T C)^{-1}C^T \eta). \tag{4.31b}
\]
and thus:

\[ B^T (A(C^T C)^{-1} A^T)^{-1} B \theta = -B^T (A(C^T C)^{-1} A^T)^{-1} A(C^T C)^{-1} C^T \eta. \]  \hspace{1cm} (4.32)

We can see that the kernel matrix is given by \( K = B^T (A(C^T C)^{-1} A^T)^{-1} B \) (this matrix is also known as the reduced Hessian matrix). Moreover, if we define \( X^T = -B^T (A(C^T C)^{-1} A^T)^{-1} A(C^T C)^{-1} C^T \), we have that \( X = -C(C^T C)^{-1} A^T (A(C^T C)^{-1} A^T)^{-1} B \) and, consequently:

\[
X^T X = B^T (A(C^T C)^{-1} A^T)^{-1} A(C^T C)^{-1} A^T (A(C^T C)^{-1} A^T)^{-1} B \\
= B^T (A(C^T C)^{-1} A^T)^{-1} A(C^T C)^{-1} A^T (A(C^T C)^{-1} A^T)^{-1} B \\
= B^T (A(C^T C)^{-1} A^T)^{-1} B.
\]

We thus have that \( V[\theta] = \sigma^2 K^{-1} \) with \( K = X^T X \) and \( X = -C(C^T C)^{-1} A^T (A(C^T C)^{-1} A^T)^{-1} B \).

Now consider the problem with regularization constraints:

\[
(\hat{w}, \hat{\theta}) \in \arg \min_{w, \theta} \frac{1}{2} (\eta - Cw)^T (\eta - Cw) \hspace{1cm} (4.33a)
\]

s.t. \( Aw + B\theta = 0 \hspace{1cm} (4.33b) \]
\[
R\theta = r. \hspace{1cm} (4.33c)
\]

The optimality conditions are given by:

\[
C^T C w + A^T \nu - C^T \eta = 0 \hspace{1cm} (4.34a)
\]
\[
B^T \nu + R^T \lambda = 0 \hspace{1cm} (4.34b)
\]
\[
Aw + B\theta = 0 \hspace{1cm} (4.34c)
\]
\[
R\theta = r. \hspace{1cm} (4.34d)
\]

In matrix form,

\[
\begin{bmatrix}
C^T C & A^T \\
B^T & R^T \\
A & B \\
R
\end{bmatrix}
\begin{bmatrix}
w \\
\theta \\
\nu \\
\lambda
\end{bmatrix}
= \begin{bmatrix}
C^T \eta \\
0 \\
0 \\
r
\end{bmatrix}. \hspace{1cm} (4.35)
\]

Using block elimination we obtain:

\[
w = (C^T C)^{-1} (-A^T \nu + C^T \nu) \hspace{1cm} (4.36a)
\]
\[
\nu = (A(C^T C)^{-1} A^T)^{-1} (B\theta + A(C^T C)^{-1} C\eta). \hspace{1cm} (4.36b)
\]

Substituting \( \nu \) into the second expression we obtain:

\[
B^T (A(C^T C)^{-1} A^T)^{-1} B\theta + B^T (A(C^T C)^{-1} A^T)^{-1} A(C^T C)^{-1} C\eta + R^T \lambda = 0 \hspace{1cm} (4.37)
\]

Now, as in the unregularized case, define \( X^T = -B^T (A(C^T C)^{-1} A^T)^{-1} A(C^T C)^{-1} C \) such that \( K = X^T X = B^T (A(C^T C)^{-1} A^T)^{-1} B \). The above expression thus becomes \( K\theta - X^T \eta + R^T \lambda = 0 \) and thus
\[ \theta = K^{-1}X^T\eta - K^{-1}R^T\lambda. \]

Substituting this in the fourth expression of the optimality conditions we obtain \[ RK^{-1}X^T\eta - RK^{-1}R^T\lambda = r \]
and thus \[ \lambda = (RK^{-1}R^T)^{-1}RK^{-1}X^T\eta - (RK^{-1}R^T)^{-1}r \]
which is equivalent to (3.12). This reveals that \( \hat{\theta} \) can be obtained (3.13) but with the new definitions of \( K \) and \( X \). From this analysis we conclude that the restriction matrix \( R = \sqrt{\Lambda_2}V_2 \) should be obtained from the eigenvalue decomposition of the kernel matrix \( K \) (the reduced Hessian matrix).

Based on our observations on the equivalence between eigenvalue constraints and PC regression, we note that a convenient way to implement this regularization approach in a more general setting is to formulate the estimation problem in a reduced space by using the reduced parameter set \( \gamma = V_1\theta \).

Specifically, we aim to solve the problem:

\[
(\hat{w}, \hat{\theta}, \hat{\gamma}) \in \arg \min_{w,\theta,\gamma} \frac{1}{2}(\eta - Cw)^T(\eta - Cw) \tag{4.38a}
\]

\[
s.t. \quad Aw + B\theta = 0 \tag{4.38b}
\]

\[
V_1^T\theta = \gamma. \tag{4.38c}
\]

where \( V_1 \) is obtained from an eigenvalue decomposition of \( K = X^TX = B^T(A(C^TC)^{-1}A^T)^{-1}B \) and an estimate of the parameters in the original space is obtained as \( \hat{\theta} = V_1\hat{\gamma} \). Simple algebraic manipulations show that, by restricting \( \theta = V_1\gamma \), we obtain a kernel matrix of the form \( K = V_1^TX^TXV_1 \) with \( X = C(C^TC)^{-1}A^T(A(C^TC)^{-1}A^T)^{-1}B \) and thus \( V[V\hat{\theta}] = \sigma^2V_1\Lambda_1^{-1}V_1 \), as desired. From this analysis we also conclude that the coefficient matrix \( V_1 \) should be obtained from the reduced hessian \( K \). We emphasize that this conclusion is both surprising and relevant because it provides a mechanism to regularize arbitrarily complex linear estimation problems (as long as the reduced Hessian is known). Moreover, the structure of the linear system (4.30) is found Newton-based optimization solvers (e.g., Ipopt) [17] and thus the proposed regularization approach can be used with such solvers.

In order to regularize the estimation problem it is necessary to obtain \( K \), which might not be directly accessible (e.g., because the estimation problem involves a complex formulation). It is well-known that it is possible to compute the covariance of the parameters by performing Monte Carlo simulations and the connection of the covariance with the kernel matrix provides a path to estimate \( K \). To illustrate the Monte Carlo approach we use the general formulation (4.28). Here, we solve a set of \( \ell = 1, ..., L \) estimation problems of the form:

\[
(w_\ell, \hat{\theta}_\ell) \in \arg \min_{w,\theta} \frac{1}{2}((\eta + \epsilon_\ell) - Cw)^T((\eta + \epsilon_\ell) - Cw) \tag{4.39a}
\]

\[
s.t. \quad Aw + B\theta = 0. \tag{4.39b}
\]

where \( \epsilon_\ell \) are samples drawn from \( \mathcal{N}(0, \sigma^2) \). From the structure of (4.32), it is clear that the parameter \( \hat{\theta}_\ell \) is simply a linear transformation of the perturbation \( (\eta + \epsilon_\ell) \) and thus we can obtain a sample
estimate of the expected value $E[\hat{\theta}]$ and covariance $\nabla[\hat{\theta}]$ as:

$$m_L = L^{-1} \sum_{\ell=1}^{L} \theta_\ell \approx E[\hat{\theta}] \quad (4.40a)$$

$$S_L = (L - 1)^{-1} \sum_{\ell=1}^{L} (\theta_\ell - m_L)(\theta_\ell - m_L)^T \approx \nabla[\hat{\theta}] \quad (4.40b)$$

from the law of large numbers we have that $\lim_{L \to \infty} m_L \to E[\hat{\theta}]$ and $\lim_{L \to \infty} S_L \to \nabla[\hat{\theta}]$. Consequently, one can use $S_L$ to estimate the kernel matrix as $K \approx \sigma^{-2} S_L$. It is also possible to extract the kernel matrix from optimization solvers such as Ipopt. This is done by solving the linear system:

$$\begin{bmatrix} C^T C & A^T \\ A & B^T \end{bmatrix} \begin{bmatrix} w \\ \theta \\lambda \end{bmatrix} = \begin{bmatrix} C^T (\eta + \epsilon) \\ 0 \\ 0 \end{bmatrix}$$

$$\quad (4.41)$$

Here, a factorization of the matrix on the left-hand side can be re-used and thus only backsolves are needed. This thus overcomes the need to solve multiple estimation problems. The computed kernel matrix $K$ is used to perform eigenvalue decomposition and to compute needed regularization quantities. The Monte Carlo approach can also be used to estimate the parameter covariance for the regularized problem. For instance, in PC regression we solve a set of $\ell = 1, ..., L$ estimation problems of the form:

$$(\hat{\omega}_\ell, \hat{\theta}_\ell, \hat{\gamma}_\ell) \in \arg\min_{w, \theta, \gamma} \frac{1}{2} \big((\eta + \epsilon)\big)^T((\eta + \epsilon) - Cw)$$

$$s.t. \quad Aw + B\theta = 0$$

$$\quad V_i^{T} \theta = \gamma.$$ 

Similarly, one can obtain the reduced Hessian for the regularized problem and the covariance matrix from solvers such as Ipopt.

## 5 Case Studies

In this section we present case studies to illustrate the concepts and to highlight insights that can be gained with the proposed framework.

### 5.1 Illustrative Example I

We apply the proposed eigenvector regularization approach to an ill-conditioned linear model in which the input data present near collinearities. A synthetic model was built with $m = 6$ parameters
and \( n = 15 \) data points. To induce collinearities, we generate input data as follows:

\[
\begin{align*}
    x_1 & \sim \mathcal{N}(0, 1) \quad (5.43a) \\
    x_2 & \sim \mathcal{N}(0, 1) \quad (5.43b) \\
    x_3 & \sim \mathcal{N}(0, 1) \quad (5.43c) \\
    x_4 & = x_1 \quad (5.43d) \\
    x_5 & = x_2 \quad (5.43e) \\
    x_6 & = x_1 + x_2. \quad (5.43f)
\end{align*}
\]

This approach dependencies between parameters \( \theta_1, \theta_4, \theta_6 \) and between parameters \( \theta_2, \theta_5, \theta_6 \) (there are a couple of clusters in this problem). No collinearities are induced by the third input.

We found that most of the variance of the inputs can be captured by the \( p = 3 \) leading eigenvalues of \( X^TX \) (the rest of the eigenvalues are close to zero). This highlights that, as expected, only three parameters can be estimated (e.g., the two clusters clusters and an additional parameter). The eigenvector matrix \( V_1 = (v_1, v_2, v_3) \) is shown in Table 2; here, we also present sparse eigenvectors obtained with elastic net and obtained with elastic net with orthogonality constraints. Figure 3 shows the parameter covariance when using unregularized estimation and PC regression. It is clear that the problem is ill-conditioned and regularization is needed to stabilize the estimates.

We compare the effect of regularization using a PC regression approach using an elastic net approach directly (denoted as EN) and using elastic net with orthogonality constraints (denoted as EN-OC). We recall that the direct use of elastic net does not provide the desired regularization effect. Here, we also explore the performance of EN and EN-OC under different values for the tuning parameters \( \kappa_1 \) and \( \kappa_2 \). Since in this case \( n > m \), we can set \( \kappa_2 = 0 \) (this does not affect the solution) [16]. Tuning parameter \( \kappa_1 \) controls the sparsity of the vector and we note that \( \kappa_1 = 0 \) is equivalent to PC regression with dense eigenvectors (\( \kappa_1 \) can affect the parameter variance). Figure 2b shows the variance as a function of \( \kappa_1 \). For EC, the variance increases as \( \tilde{V}_1 \) becomes sparser (the numbers in parenthesis are the number of nonzeros in each eigenvector) but the quality of the eigenvectors degrades (the variance increases). For a large value of \( \kappa_1 \) we observe that all eigenvectors only contain one entry (denoted as \( (1, 1, 1) \)). Figure 2b also shows that EN-OC keeps the parameter variance at the minimum (optimal) value obtained with dense PC regression (the variance does not increase with \( \kappa_1 \) but the sparsity does increase). This result is surprising, as it indicates that one can improve sparsity without compromising optimality. However, we also observe that EN-OC cannot fully sparsify the eigenvectors (in the limit we obtain eigenvectors with \( (3, 6, 1) \) nonzero entries); this is because the orthogonality constraints only allow spanning a limited subspace of \( X \).

The dense eigenvectors \( V_1 \) and sparse eigenvectors \( \tilde{V}_1 \) calculated with EN and EN-OC are presented in Table 2. The value of \( \kappa_1 \) for EN was chosen based on the trade-off between variance and sparsity while for EN-OC we use a limiting value at which sparsity is no longer affected. Interestingly, we observe that the sparsity structure of the eigenvectors of EN-OC reveals the clusters of parameters associated with the dependent columns of the input matrix (parameters \( \theta_1, \theta_4, \theta_6 \), parameters \( \theta_2, \theta_5, \theta_6 \), and parameter \( \theta_3 \). This illustrates how sparse eigenvectors can help reveal parameter clusters. We also observe that PC regression and EN do not identify these clusters.
Table 1: Eigenvectors and corresponding eigenvalues of $X^T X$ (Example I).

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.502</td>
<td>-0.138</td>
<td>-0.224</td>
<td>-0.502</td>
<td>-0.138</td>
<td>-0.639</td>
</tr>
<tr>
<td></td>
<td>-0.253</td>
<td>0.554</td>
<td>-0.409</td>
<td>-0.252</td>
<td>0.554</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>0.244</td>
<td>-0.222</td>
<td>-0.884</td>
<td>0.244</td>
<td>-0.222</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>0.051</td>
<td>-0.346</td>
<td>0.0</td>
<td>-0.661</td>
<td>-0.264</td>
<td>0.610</td>
</tr>
<tr>
<td></td>
<td>0.587</td>
<td>0.324</td>
<td>0.0</td>
<td>0.298</td>
<td>-0.614</td>
<td>0.290</td>
</tr>
<tr>
<td></td>
<td>0.527</td>
<td>0.633</td>
<td>0.0</td>
<td>-0.316</td>
<td>-0.422</td>
<td>-0.211</td>
</tr>
</tbody>
</table>

| $\Lambda$ | 48.8 | 31.4 | 12.3 | 1.6e-5 | 5.8e-6 | 1.9e-6 |

Figure 2a shows the Frobenius norm of the covariance matrix obtained under parameter subset selection (for all possible combinations with $p = 3$ fixed parameters) and shows the norm obtained with PC regression. We can see that none of the subset selections reach the minimum variance of PC regression (corroborating that PC regression is optimal and that subset selection is inherently suboptimal). Interestingly, we have also have found that the variance of the sparsest eigenvectors found with EN (one nonzero entry per eigenvector) is similar to that of the best subset selection. This indicates that one can find a suitable subset by using EN (thus avoiding a combinatorial search).

Figure 1: Covariance matrix for estimated parameters $V[\hat{\theta}]$ (Example I).

5.2 Illustrative Example II

We now analyze a simple setting with more parameters than data points ($m > n$). Since in this case the kernel matrix $K = X^T X$ is singular, unregularized estimation cannot be directly used. Here, we compare results using ridge regularization and PC regression with dense and sparse eigenvectors. The results presented use a random matrix $X \in \mathbb{R}^{n \times m}$ with $m = 8$, $n = 5$ and model $\eta = X\theta + e$, with $e \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma^2 = 1 \times 10^{-4}$. 

14
Table 2: Eigenvectors $V_1$ and sparse eigenvectors $\tilde{V}_1$ obtained with elastic net (EN) and elastic net with orthogonality constraints (EN-OC) (Example I).

<table>
<thead>
<tr>
<th></th>
<th>$V_1$</th>
<th>$\tilde{V}_1$ (EN)</th>
<th>$\tilde{V}_1$ (EN-OC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$v_1$</td>
<td>$\tilde{v}_1$</td>
<td>$\tilde{v}_1$</td>
</tr>
<tr>
<td>$X_1$</td>
<td>-0.502</td>
<td>0</td>
<td>-0.577</td>
</tr>
<tr>
<td>$X_2$</td>
<td>-0.138</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$X_3$</td>
<td>-0.224</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$X_4$</td>
<td>-0.502</td>
<td>0</td>
<td>-1.0</td>
</tr>
<tr>
<td>$X_5$</td>
<td>-0.138</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$X_6$</td>
<td>-0.639</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3: Eigenvectors and corresponding eigenvalues of $X^T X$ (Example II).

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.462</td>
<td>0.246</td>
<td>-0.371</td>
<td>0.441</td>
<td>0.468</td>
<td>0.350</td>
<td>-0.116</td>
<td>-0.198</td>
</tr>
<tr>
<td></td>
<td>0.355</td>
<td>0.675</td>
<td>0.412</td>
<td>-0.319</td>
<td>-0.178</td>
<td>0.258</td>
<td>0.075</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>-0.201</td>
<td>0.056</td>
<td>-0.123</td>
<td>-0.066</td>
<td>-0.795</td>
<td>0.045</td>
<td>-0.299</td>
</tr>
<tr>
<td></td>
<td>0.106</td>
<td>0.350</td>
<td>0.564</td>
<td>-0.482</td>
<td>0.333</td>
<td>0.070</td>
<td>0.075</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>-0.484</td>
<td>-0.242</td>
<td>-0.545</td>
<td>-0.319</td>
<td>-0.221</td>
<td>-0.153</td>
<td>-0.045</td>
<td>-0.299</td>
</tr>
<tr>
<td></td>
<td>0.621</td>
<td>0.277</td>
<td>-0.479</td>
<td>-0.389</td>
<td>-0.333</td>
<td>0.305</td>
<td>0.203</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>-0.115</td>
<td>0.125</td>
<td>-0.354</td>
<td>0.082</td>
<td>0.0904</td>
<td>-0.136</td>
<td>0.052</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>0.131</td>
<td>0.443</td>
<td>-0.011</td>
<td>0.443</td>
<td>-0.619</td>
<td>-0.198</td>
<td>-0.590</td>
<td>-0.291</td>
</tr>
</tbody>
</table>

| $\Lambda$ | 19.0 | 16.7 | 5.8 | 3.9 | 0.5 | 0 | 0 | 0 |

We have found that sparse and dense PC regression can delivers parameter estimates with a SSE of zero (regardless of the sparsity level used). We again found that orthogonality constraints are essential to control the variance of the parameters. In particular, we found that using orthogonality constraints maintains a minimum variance regardless of the number of nonzeros in the eigenvector matrix (see Figure 3a). Interestingly, we found that ridge regularization can only achieve the same level of variance as PC regression by sacrificing SSE (see Figure 3b). This illustrate that PC regression offers different regularization behavior than objective regularization.

5.3 Enzymatic Reactions

Enzymatic reactions can be represented by elementary steps and described by the mass action kinetics. A simple mechanism with one substrate $S$ and one enzyme $E$ resulting in one product $P$ can be expressed as

$$ E + S \xrightarrow{k_1} ES \xrightarrow{k_2} EP \xrightarrow{k_3} E + P. $$

Suppose, for example, that the rate of reaction of the first step in both directions is many orders of magnitude higher than the following steps. This implies that the substrate-enzyme binding and dissociation step is considerably faster, which is said to be in a quasi-equilibrium state. Michaelis-Menten
Figure 2: (a) Frobenius norm of parameter covariance matrix obtained with all possible subset selections with $p = 3$ (circles) and obtained with PC regression (red line). (b) Frobenius norm of the covariance matrix obtained with sparse PC regression with elastic net (EN) and elastic net with orthogonality constraints (EN-OC). The sparsest eigenvectors (with one nonzero entry) is equivalent to fixing parameters $\theta_1, \theta_4$ and $\theta_5$.

Figure 3: (a) Frobenius norm of the covariance matrix of the estimated parameters calculated with sparse PC regression (both approaches) as a function of the number of nonzero entries in each eigenvector in $\tilde{V}_1$. (b) Frobenius norm of the covariance matrix of the estimated parameters (circles) and sum of the squared errors of $\hat{\eta}$ (stars) using Ridge regression as a function of its parameter (Example II).

is the most used model for describing enzymatic reactions and is derived under a quasi-equilibrium assumption. Using the same principles, other expressions have been developed to account for additional conditions, such as inhibition and activation. Using this types of equations require knowledge
Table 4: Complete set of eigenvectors $V_1$ and $V_2$ and sparse eigenvectors $\tilde{V}_1$ obtained with elastic net with orthogonality constraints (EN-OC).

<table>
<thead>
<tr>
<th></th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$\tilde{v}_1$</th>
<th>$\tilde{v}_2$</th>
<th>$\tilde{v}_3$</th>
<th>$\tilde{v}_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>0.102</td>
<td>-0.587</td>
<td>-0.224</td>
<td>0.066</td>
<td>0.768</td>
<td>0</td>
<td>-0.640</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$k_{-1}$</td>
<td>-0.122</td>
<td>0.705</td>
<td>0.268</td>
<td>-0.079</td>
<td>0.640</td>
<td>0</td>
<td>0.768</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$k_2$</td>
<td>0.073</td>
<td>0.377</td>
<td>-0.905</td>
<td>0.181</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>$k_3$</td>
<td>-0.755</td>
<td>-0.073</td>
<td>0.039</td>
<td>0.651</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$k_{-3}$</td>
<td>0.632</td>
<td>0.100</td>
<td>0.238</td>
<td>0.730</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>4.335</td>
<td>1.921</td>
<td>0.495</td>
<td>0.033</td>
<td>2.2e-16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

about the enzymatic reactions being modeled and manual selection of the appropriate equation describing such reactions. In this case study, we wish to apply the sparse PC regularization approach to identify steps in quasi-equilibrium and estimate kinetic constants using the complete mass action description and concentration data without further assumptions.

Consider the simple enzymatic reaction just described in an open steady-state system:

\[
-k_1 x_E x_S + k_{-1} x_{ES} + F(x_{in}^S - x_S) = 0 \quad (5.44a)
\]
\[
-k_1 x_E x_S + k_{-1} x_{ES} + k_3 x_{EP} - k_{-3} x_{EP} + F(x_{in}^E - x_E) = 0 \quad (5.44b)
\]
\[
k_1 x_E x_S - k_{-1} x_{ES} - k_2 x_{ES} + F(x_{in}^{ES} - x_{ES}) = 0 \quad (5.44c)
\]
\[
k_2 x_{ES} - k_3 x_{EP} + k_{-3} x_{EP} + F(x_{in}^{EP} - x_{EP}) = 0 \quad (5.44d)
\]
\[
k_3 x_{EP} - k_{-3} x_{EP} + F(x_{in}^{EP} - x_P) = 0 \quad (5.44e)
\]

where $k_j$ is the kinetic constant of the forward reaction of the $j^{th}$ step, $k_{-j}$ is the kinetic constant of the backward reaction of the $j^{th}$ step, $x_i$ is the volume concentration of the $i^{th}$ specie, $x_{in}^i$ is the concentration of the $i^{th}$ specie in the inlet and $F$ is the volume flow rate divided by the volume of the system. It is important to note that steady state data does not provide enough information for individually identifying steps with considerably larger kinetic constants that operate in a different timescale. If we assume that all species are measured and $F$ and $x_{in}^i$ are known, we have a linear model with respect to the parameters. We generated $X$ with 10 data points by varying $F$ from 0.01 to 2.5, fixing the inlet flow to contain only enzyme and substrate in equal concentrations, and calculating all species concentration at steady state. $k_1$ and $k_{-1}$ were set to be 9 orders of magnitude larger than $k_2$, $k_3$ and $k_{-3}$, which creates a quasi-equilibrium for the first step.

The eigenvalues and eigenvectors of the kernel matrix are shown in Table 4. We can see that $K$ is ill-conditioned and the coefficients of the covariance matrix corresponding to $k_1$ and $k_{-1}$ are large (Figure 4). We have found that PC regression and EN-OC sparsification provide the same reduced variance and estimates. However, EN-OC (in its sparsest form) has the advantage of explicitly identifying steps in quasi-equilibrium, as shown by the second sparse eigenvector (Table 4).

We also applied this approach for the general modifier mechanism of Botts and Morales [18]:

17
Depending on the values of the parameters, species $M$ can act as an inhibitor or as an activator. Usually, quasi-equilibrium is assumed for the reversible reactions so that Michaelis-Menten type of equations can be used. Our goal is to use concentration data for all species to identify quasi-equilibrium steps and the actual role of $M$. For non-competitive linear inhibition with zero rate of reaction for steps 1', 2', 3', 4' and 7, and kinetic constants for steps 1, 5 and 6 much larger than for steps 2, 4 and 3, the sparse eigenvectors of EN-OC identify steps 1, 1', 5, and 6 to be in quasi-equilibrium (see Table 6). In contrast, PC regression correctly estimates reaction steps with zero rate but does not reveal quasi-equilibrium states (see Table 5). Other configurations were also tested (such as cases with essential activation and competitive inhibition) and we consistently found that EN-OC correctly identified steps that were in a different timescale and those with a zero rate of reaction. This example illustrates...
Table 5: Dense eigenvectors $V_1$ for Botts-Morales example.

<table>
<thead>
<tr>
<th></th>
<th>$k_1$</th>
<th>$k_{-1}$</th>
<th>$k_2$</th>
<th>$k_3$</th>
<th>$k_{-3}$</th>
<th>$k_4$</th>
<th>$k_1'$</th>
<th>$k_{-1}'$</th>
<th>$k_2'$</th>
<th>$k_3'$</th>
<th>$k_{-3}'$</th>
<th>$k_4'$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.153</td>
<td>-0.153</td>
<td>-0.011</td>
<td>-0.061</td>
<td>0.029</td>
<td>-0.0</td>
<td>0.074</td>
<td>0.074</td>
<td>0.0</td>
<td>0.0</td>
<td>0.002</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>-0.595</td>
<td>0.595</td>
<td>0.3</td>
<td>0.036</td>
<td>-0.012</td>
<td>-0.017</td>
<td>0.18</td>
<td>0.18</td>
<td>-0.274</td>
<td>0.0</td>
<td>-0.003</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.082</td>
<td>-0.082</td>
<td>0.215</td>
<td>-0.036</td>
<td>0.013</td>
<td>-0.025</td>
<td>0.492</td>
<td>-0.492</td>
<td>-0.274</td>
<td>0.0</td>
<td>-0.052</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>-0.032</td>
<td>0.032</td>
<td>-0.207</td>
<td>-0.036</td>
<td>-0.127</td>
<td>0.065</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.246</td>
<td>0.0</td>
<td>-0.052</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.131</td>
<td>-0.131</td>
<td>0.743</td>
<td>0.238</td>
<td>0.123</td>
<td>-0.281</td>
<td>-0.075</td>
<td>-0.075</td>
<td>0.034</td>
<td>0.0</td>
<td>0.005</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>-0.064</td>
<td>0.064</td>
<td>-0.222</td>
<td>-0.317</td>
<td>0.174</td>
<td>-0.075</td>
<td>0.925</td>
<td>0.925</td>
<td>-0.802</td>
<td>0.0</td>
<td>-0.132</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.075</td>
<td>0.075</td>
<td>0.305</td>
<td>-0.295</td>
<td>-0.053</td>
<td>0.0</td>
<td>0.248</td>
<td>0.248</td>
<td>-0.132</td>
<td>0.0</td>
<td>-0.131</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.208</td>
<td>0.208</td>
<td>0.255</td>
<td>0.001</td>
<td>-0.376</td>
<td>-0.227</td>
<td>-0.16</td>
<td>-0.16</td>
<td>0.071</td>
<td>0.0</td>
<td>-0.256</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>-0.205</td>
<td>-0.205</td>
<td>0.173</td>
<td>0.604</td>
<td>-0.085</td>
<td>-0.227</td>
<td>0.048</td>
<td>0.048</td>
<td>0.248</td>
<td>0.0</td>
<td>0.003</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.115</td>
<td>0.401</td>
<td>0.744</td>
<td>-0.006</td>
<td>0.036</td>
<td>0.036</td>
<td>0.071</td>
<td>0.0</td>
<td>0.067</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.045</td>
<td>0.045</td>
<td>0.14</td>
<td>0.262</td>
<td>0.482</td>
<td>0.016</td>
<td>0.365</td>
<td>0.365</td>
<td>0.248</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

how the use of sparsification techniques can help identify relevant clusters of parameters.

6 Conclusions and Future Work

We have explored strategies to regularize ill-posed parameter estimation problems by using constraints. We show how to construct optimal constraints that minimize parameter covariance by exploiting information from the reduced Hessian matrix. We also derive a elastic net strategy to sparsify the constraints and facilitate their interpretability. We show that this approach can identify clusters of parameters in an effective manner and we discuss how to apply it in a general setting. Our analysis also highlights connections with parameter subset selection. As part of future work, we are interested in applying the proposed regularization approach to nonlinear estimation problems. This can be done by using regularization insider an optimization solvers such as Ipopt. The effect of this approach could be intriguing because such solvers commonly use Hessian regulation (similar to objective regularization) and this can lead to search steps of poor quality. Our aim is thus to use eigenvector information to enhance the quality of the search steps.
Table 6: Sparse eigenvectors $\tilde{V}_1$ or Botts-Morales example calculated with elastic net with orthogonality constraint.

<table>
<thead>
<tr>
<th></th>
<th>$\tilde{V}_1$ (EN-OC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_{-1}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_2$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_{-3}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_3$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_{-3}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_4$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_{-4}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_5$</td>
<td>0.707</td>
</tr>
<tr>
<td>$k_{-5}$</td>
<td>-0.707</td>
</tr>
<tr>
<td>$k_6$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_{-6}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_7$</td>
<td>0.0</td>
</tr>
<tr>
<td>$k_{-7}$</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References


