Inverse problem approach to regularized regression models with application to predicting recovery after stroke

Youssef Hbid, Khaladi Mohamed, Charles D.A. Wolfe, Abdel Douiri

Received: 23 September 2019 | Revised: 27 June 2020 | Accepted: 2 August 2020

DOI: 10.1002/bimj.201900283

Biometrical Journal

RESEARCH PAPER

Inverse problem approach to regularized regression models with application to predicting recovery after stroke

Youssef Hbid[1,2,3] | **Khaladi Mohamed**[1,2] | **Charles D.A. Wolfe**[4,5] | **Abdel Douiri**[4,5]

[1] LMDP, Cadi Ayyad University, Marrakech, Morocco

[2] UMMISCO, IRD, France

[3] Laboratoire Jacques-Louis Lions, Sorbonne University, Paris, France

[4] School of Population Health and Environmental Sciences, King's College London, London, United Kingdom

[5] National Institute for Health Research Biomedical Research Centre, Guy's and St Thomas' NHS Foundation Trust and King's College London, London, United Kingdom

Correspondence
Abdel Douiri, King's College London, School of Population Health and Environmental Sciences, London, United Kingdom.
Email: abdel.douiri@kcl.ac.uk

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Regression modelling is a powerful statistical tool often used in biomedical and clinical research. It could be formulated as an inverse problem that measures the discrepancy between the target outcome and the data produced by representation of the modelled predictors. This approach could simultaneously perform variable selection and coefficient estimation. We focus particularly on a linear regression issue, $Y \sim N(X\beta, \sigma I_n)$, where $\beta \in \mathbb{R}^p$ is the parameter of interest and its components are the regression coefficients. The inverse problem finds an estimate for the parameter β, which is mapped by the linear operator $(L : \beta \longrightarrow X\beta)$ to the observed outcome data $Y = X\beta + \epsilon$. This problem could be conveyed by finding a solution in the affine subspace $L^{-1}(Y)$. However, in the presence of collinearity, high-dimensional data and high conditioning number of the related covariance matrix, the solution may not be unique, so the introduction of prior information to reduce the subset $L^{-1}(Y)$ and regularize the inverse problem is needed. Informed by Huber's robust statistics framework, we propose an optimal regularizer to the regression problem. We compare results of the proposed method and other penalized regression regularization methods: ridge, lasso, adaptive-lasso and elastic-net under different strong hypothesis such as high conditioning number of the covariance matrix and high error amplitude, on both simulated and real data from the South London Stroke Register. The proposed approach can be extended to mixed regression models. Our inverse problem framework coupled with robust statistics methodology offer new insights in statistical regression and learning. It could open a new research development for model fitting and learning.

KEYWORDS
regression modelling, regularization, robust statistics, statistical inverse problem, stroke

1 | INTRODUCTION

Advanced regression modelling was refined by the work of Fisher. Fisher combined the work of Gauss and Pearson to develop a theory of the properties of least squares estimation (Aldrich, 2005). Owing to Fisher's work,

regression analysis is not just used for fitting independent and dependent variables, but to identify causal relationships between predictors and outcome to contribute to the evidence on causal relationships between presumed causes or exposure and an observed effect in terms of Bradford Hill Criteria (Bradford Hill, 1965). Since Fisher's theory was published, there have been several developments in regression analysis such as parametric and nonparametric regression (Pagan & Ullah, 1999), Bayesian regression (Broemeling, 1985) and regularized regression (Bühlmann & van de Geer, 2011). Nowadays, using multivariate regression is an essential tool often used in clinical trials and epidemiological studies.

Gauss introduced the normal distribution as the error distribution for which ordinary least square or maximum likelihood (OLS/ML) is optimal (Aitken, 1936; Gauss & Davis, 1857). However, real-world data do not often satisfy these classical assumptions (Baltagi, 2008). OLS- or ML-based methods may fail to predict or to give meaningful interpretations in certain scenarios. To overcome these limitations, many techniques have been developed such as ridge regression (Hoerl & Kennard, 1970), lasso (Tibshirani, 1994), adaptive-lasso (Zou, 2006) and elastic-net (Zou & Hastie, 2005).

From a mathematical perspective, the regression fitting problem could be formulated as a minimization of an inverse problem that measures the discrepancy between the observed outcome and the data produced by representation of the modelled predictors. The inverse problem approach could simultaneously perform variable selection and coefficient estimation.

In particularly, on the general linear regression problem (McCullagh & Nelder, 1989), we assume that:

$X \in M_{n,p}(\mathbb{R})$, given $Y \in \mathbb{R}^n$ find $\beta \in \mathbb{R}^p$ the parameter of interest such that:

$g(Y) = X_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p$, where $X_1, X_2, \ldots, X_p$ columns of X are a set of real valued integrable random variables. In statistics we say that observation could be explained by $X_1, X_2, \ldots, X_p$ covariates.

Since the measurements are always accompanied by measurement errors and in some situations jump discontinuities, the minimization of the distance between observed and modelled data will give unsatisfactory solutions, therefore some prior information is needed in order to regularize the solution.

The problem of minimization could be stated as follows:

$$\begin{cases} \underset{\beta\in\mathbb{R}^p}{\text{Minimize}} & -log\left(\sum_{i=1}^{n} p(y_i|x_i,\beta,\sigma)\right) + \lambda \sum_{i=1}^{p} \psi(\beta_i) \\ \text{subject to} & \beta \in C \end{cases} \tag{1}$$

- $\lambda > 0$ is the parameter of regularization that adjust the trade-off between the proximity of observations and the degree of regularity.
- C is a constraint set of the prior knowledge on β.
- ψ: $\mathbb{R} \to \mathbb{R}$ is a regularization function.

The fidelity term $-log\left(\sum_{i=1}^{n} p(y_i|x_i,\beta,\sigma)\right)$ can be regarded as a tolerant expression of the constraint equation.

Note that this could be expressed considering least square (LS) as a particular case of (1):

$$\begin{cases} \underset{\beta\in\mathbb{R}^p}{\text{Minimize}} & \frac{1}{2}||Y - X\beta||^2 + \lambda \sum_{i=1}^{p} \psi(\beta_i) \\ \text{subject to} & \beta \in C, \end{cases} \tag{2}$$

and is expressed in the integral form as:

$$\begin{cases} \underset{\beta\in\mathbb{R}^p}{\text{Minimize}} & \frac{1}{2}||Y - X\beta||^2 + \lambda \int \psi(\beta)\, d\beta \\ \text{subject to} & \beta \in C, \end{cases} \tag{3}$$

where ψ: $\mathbb{R}^p \to \mathbb{R}$.

Choosing the most adequate prior function for a given inverse problem is not simple, and need to take into account two important considerations in practice: the accuracy (error) of prediction on future data and meaningful interpretations of the model. To illustrate the error linked to the regression fitting problem, we investigate in a simple example (bellow)

which shows the impact of imperfect observation on the accuracy of estimate.

$$X = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \quad Y = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \beta = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad Y + \delta Y = \begin{pmatrix} 32.1 \\ 23.9 \\ 33.1 \\ 31.9 \end{pmatrix} \quad \beta + \delta\beta = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}$$

We aim to revisit regularized regression modelling using an inverse problem framework, which represents a novel way to advance the methodologies for statistical problem. We investigate the statistical inverse problem in the context of regression modelling, explaining the importance of regularization in analytical predictive models. The theoretical analysis is supported by a simulation study in order to evaluate different regularization strategies in low- and high-dimensional settings. Building upon Huber's robust statistics (Huber, 1981) and inverse problem framework, we characterize and propose a new regularization function for the regression problem. We evaluate the performance of the proposed method and other existent methods (lasso, adaptive-lasso, ridge, elastic-net) using simulations and real data from South London Stroke Register (SLSR).

2 | METHODOLOGY

2.1 | Statistical inverse problem framework in regression context

Inverse problem consists in finding an estimate for $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ which is mapped by an operator L: $\beta \longrightarrow X\beta$ to fit data Y (Vogel, 2002).

The measured data $\bar{Y}$ approximate the perfect data Y within error estimate:

$||\bar{Y} - Y|| \leq \delta$, where $\delta > 0$.

Referring to Hadamard's definition for inverse problem (Hadamard, 1902), we say that the problem is ill-posed if the injectivity, surjectivity and stability cannot be assured.

We will study the influence of the regularity term such that the problem has a unique, stable and statistically meaningful and interpretable solutions.

In the regression modelling, we suppose that we observe an outcome $Y \in \mathbb{R}^n$ and a predictor Matrix X $\in M_{n,p}(\mathbb{R})$, whose columns $X_1, \dots, X_p \in \mathbb{R}^n$ correspond to predictor variables.

Using Hadamard's definition in the context of regression modelling, we will establish the surjectivity, the injectivity through the selection of a meaningful solution by introducing a prior information, and finally formulate a generalization of the regularized inverse problem.

To formulate a generalization of the regularized regression problem, we use a heuristic analysis based on four cases:

The simple case is when X is bijective ($n = p$), the problem is analytically well-posed, the solution of $Y = X\beta$ may verify the Hadamard conditions and is unique, but still be very sensitive to small perturbations. Also, when X is ill-conditioned, $X^{-1}Y$ amplify the error which yields to an irregular solution.

However, when X is not surjective ($Ran(X) < n \Leftrightarrow p < n$), the observation Y may not belong to $Ran(X)$, that is, Y $\notin$ $Ran(X)$. Then it seems natural to replace Y by it's projection onto $Ran(X)$, that is, $\bar{Y} = \underset{Ran(X)}{\text{Proj}}$ Y to re-establish the surjectivity. The problem becomes $\bar{\beta} = argmin||Y - X\beta||^2$, where $\bar{\beta}$ is the unique solution of the normal equation $X^tY = X^tX\beta$ called least square solution.

Moreover, when X is not injective ($Ran(X) < p$), we have an infinity of solutions forming an affine variety of dimension ($n - p$), that is, all vectors in the affine subspace $X^{-1}\{Y\}$ are solutions, hence we can select a meaningful solution in the subspace $X^{-1}\{Y\}$ by introducing a priori information which involves the choice of some inference principle such as smoothness of the solution or other restoration criteria to reduce the subspace $X^{-1}\{Y\}$ and regularize the inverse problem.

The problem becomes:

$$\begin{cases} \underset{\beta \in \mathbb{R}^p}{\text{Minimize}} & \Psi(\beta) \\ \text{subject to} & \beta \in X^{-1}\{Y\}, \end{cases} \tag{4}$$

where $\Psi(\beta)$ could be, for example,

$\Psi(\beta) = \sum_{j=1}^{p} |\beta_j|$ (Lasso),

$\Psi(\beta) = \sum_{j=1}^{p} w_j|\beta_j|$, where $w_j = \frac{1}{|\hat{\beta}_j|^v}$ and $v > 0$. w_j are the adaptive weights. As suggested by Zou (2006), $\hat{\beta}_j$ could be the OLS estimates or l_2 penalized estimator in high-dimensional data (adaptive-lasso).

$\Psi(\beta) = \sum_{j=1}^{p} \beta_j^2$ (ridge)

$\Psi(\beta) = \sum_{j=1}^{p} |\beta_j - \beta_{j-1}|$ (total variation)

$\Psi(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$ (elastic-net)

Furthermore, when X is neither injective nor surjective, the combination of these strategies for solving the inverse problem leads to recover a solution that optimizes a prior function $\Psi(\beta)$ over the the affine subspace $X^{-1}\{\bar{Y}\}$, where $\bar{Y} = \underset{Ran(X)}{\text{Proj}}\, Y$.

This can be expressed as:

$$\begin{cases} \underset{\beta \in \mathbb{R}^p}{\text{Minimize}} & \Psi(\beta) \\ \text{subject to} & \beta \in X^{-1}\{\bar{Y}\}, \end{cases} \tag{5}$$

where $\Psi : F(\mathbb{R}, \mathbb{R}^p) \longrightarrow \mathbb{R}$ and $F(\mathbb{R}, \mathbb{R}^p)$ is the space of functions defined from $\mathbb{R}$ to $\mathbb{R}^p$.

A standard and equivalent strategy of the problem (5) is to reformulate the problem in a minimization of the fidelity term $||Y - X\beta||$ plus a prior term $\Psi(\beta)$. Using Lagrange multiplier, this could be expressed as problem (3).

From inverse problem perspective, we aim to find a regularization function Ψ and a scale λ that optimize problem (3).

2.2 | Link with the Bayesian framework

The Bayesian approach consist on finding the most probable β knowing the measured data Y and the forward problem model $X\beta$.

P(β) represent the prior probability of the model object β (coefficient) and $P(Y|\beta)$ is the measurement model probability (likelihood) of generating a measure Y given an ideal model object β.

Bayes theorem (Broemeling, 1985) defined as: $P(\beta|Y) = P(Y|\beta)P(Y)/P(Y)$ express the posterior probability $P(\beta|Y)$ of a model given the object. $P(Y)$ is merely a normalization constant once the measures Y are given.

The maximum a posterior (MAP) allows to formalize inverse problem as a statistical problem, where the posterior probability $P(\beta|Y)$ has to be maximized with respect to β. By taking the logarithm, we can see that we have to minimize $E(\beta) = -logP(\beta) - logP(Y|\beta)$.

To convert the problem (3) into a prior distribution over expected object β we use a Boltzmann or Gibbs distribution of the form: $P(\beta) = \frac{1}{Z}\, exp(-\int \psi(\beta) d\beta)$ (Geman, 1988). This prior model is then combined with the model based on the measurements with a specific noise: $P(Y|\beta) = \frac{1}{Z'} exp\left(-\frac{1}{\lambda}||Y - X\beta||^2\right)$, where Z and Z' are normalization constants, called also, the partitions functions. Hence the regularized Bayesian model is expressed as the problem (3).

If we assume that we have a Gaussian noise model, we presume that $||.|| = ||.||_2$ (Least square).

2.3 | Proposed method: New hybrid regularization function

We consider the problem (5) and we define:

$$\Psi(\beta) = \int_{\Omega} \psi(|\beta(t)|)\, \mathrm{d}t, \tag{6}$$

where

$$\Omega \subset \mathbb{R} \mapsto \mathbb{R}^p \mapsto \mathbb{R}$$

$$t \mapsto \beta(t) \mapsto \psi(|\beta(t)|).$$

To ensure that $\int_\Omega \psi(|\beta(t)|)\,\mathrm{d}t$ is convex, we need the following hypothesis:

i) Ω is a convex set of $\mathbb{R}$
ii) ψ is a twice differentiable piece-wise function in $\mathbb{R}$
iii) $\psi(0) = 0; \psi'(0) = 0$
iv) $\psi''(s) \geq 0$ and $\psi'(s) \geq 0$ for all $s \geq 0$

Recall of the problem:
The problem consists to find the optimal minimizer of the following functional over a convex domain $\Omega \subset \mathbb{R}$:

$$\underset{\beta}{\text{Minimize}}(\beta) = \frac{1}{2}||Y(t) - X\beta(t)||^2 + \lambda \int_\Omega \psi(|\beta(t)|)\,\mathrm{d}t. \tag{7}$$

Suppose that $E(\beta)$ has a minimum point $\widehat{\beta}$, then the problem satisfies the following proposition.

Proposition 2.1. *$E'(\beta)$ satisfies the following equation:*

$$X^t(X\beta(t) - Y(t)) + \lambda \frac{\psi'(|\beta(t)|)}{|\beta(t)|}|\beta(t)| = 0. \tag{8}$$

Proof 2.1. See Appendix A1.

Browning idea from robust statistics, we propose the Huber function as a choice of the ψ function (regularizer) which is optimal and verify the general hypothesis:

$$\psi_\sigma(|\beta|) = \begin{cases} \frac{1}{2}\beta^2 & if\ |\beta| \leq \sigma;\ \sigma > 0 \\ \sigma\left(|\beta| - \frac{\sigma}{2}\right) & otherwise. \end{cases} \tag{9}$$

The subgradient of ψ is given by

$$\frac{\psi'_\sigma(|\beta|)}{|\beta|} = \begin{cases} \frac{\beta}{|\beta|} & if\ |\beta| \leq \sigma \\ \frac{\sigma sign(\beta)}{|\beta|} & otherwise. \end{cases} \tag{10}$$

The Huber norm is often used as a loss function and is shown to be robust and competitive against the squared loss. In our study, the Huber function is used as a regularization term. This approach is frequently used in optical tomography (Douiri, Schweiger, Riley & Arridge, 2005) and image resolution (Unger, Pock, Werlberger, & Bischof, 2010) and has shown its efficiency.

Huber regularizer enjoys the properties of L_1 and L_2 norms on the model parameters. The proposed regularizer is quadratic near the origin and linear far away from the origin which penalizes the outliers less severely. To this end, σ was specified to ensure a smooth link between quadratic to linear. As recommended by Huber, the constant σ regulates the amount of robustness which is in the range between 1 and 2, commonly chosen $\sigma = 1.5$. Note that σ could be adaptive or chosen by L-curve method (Hansen, 2001).

To solve problem (7) using the proposed regularizer and other regularization methods cited above, we used a convex optimization–based methodology using (CVXR package) (Fu, Narasimhan, & Boyd, 2017).

2.4 | Collinearity, conditioning of covariance matrices and Belsley, Kuh and Welsch test

Belsley, Kuh and Welsch test (Belsley, Kuh, & Welsch, 1980) is a method that classifies data affected by collinearity using the conditioning number of the related covariance matrix. We note C_i the conditioning index of the covariance matrix, where $C_i = d_i/d_{min}$ and $d_i = +\sqrt{\lambda_i}$ are the singular values of the covariance matrix.

We sort in descending order d_i's such that $d_k = d_{max} > d_{k-1} > ... > d_2 > d_1 = d_{min}$, then $d_1/d_{min} < d_2/d_{min} < ... < d_{k-1}/d_{min} < d_{max}/d_{min}$ and therefore, $C_1 < C_2 < C_3 < ... < C_{k-1} < C_k$, where C_k is the conditioning number of the covariance matrix.

Using simulations, Belsley, Kuh and Welsch showed that:

$$\begin{cases} \text{If} & 30 < C_i < 100 & \text{a collinearity exists} \\ \text{If} & C_i > 100 & \text{a strong collinearity exists.} \end{cases}$$

In health data, the conditioning number of the covariance matrix is often larger than 100. This test could give a measure of collinearity and provides related number of collinearity relationship.

3 | SIMULATIONS

The theoretical framework was applied by simulation to perform model selection and validation in regression models. We investigated five methods for penalized linear regression; Ridge, lasso, adaptive-lasso, elastic net and the proposed method (Huber regularizer). The statistical analysis was performed using *R statistical software*. The simulated data consisted of a range of independent data in low- ($n > p$) and high-dimensional ($p >> n$) sets, under different error amplitudes and a high condition number (strong collinearity). Each data set was divided into training and test sets. The five methods were used to fit optimal models in each of the training sets. The fitted models were used to assess the performance of predictions in the corresponding validation data sets. To assess the accuracy, we compute the mean square error (MSE), standard error and extract the number of nonzero $\hat{\beta}$-coefficients. The procedure was repeated 50 times per example. All predictors variables X were continuous multivariate normal distributed. The predictors were generated by sampling from a multivariate normal distribution with the following probability density function:

$p_X(x) = (\frac{1}{2\pi})^{\frac{n}{2}} . \frac{1}{\sqrt{det(\Sigma)}} . exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$ where μ is the mean vector and Σ is the covariance matrix. For all x we set $\mu = 0$ and $\text{Var}(x) = 1$. Through a singular value decomposition of X (SVD), we set the condition number of X. Each predictor variable was assigned a predetermined β-value. Therefore, we obtained a β vector that consisted of the corresponding β-values. In all examples, the simulated data set was divided into three partitions, two for training set and one for validation (test) set in each data set.

Case $n > p$:

- In example 1, we simulate 50 data sets, each with $n = 20$ from $y = X\beta + \sigma\epsilon$ where $\epsilon \sim N(0, 1)$. We set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $\sigma = 3$ and $cond(X) = 100$ (strong collinearity exist).
- In example 2, we simulate 50 data sets, each with $n = 20$ from $y = X\beta + \sigma\epsilon$ where $\epsilon \sim N(0, 1)$. We set $\beta = (0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80)$, $\sigma = 3$ and $cond(X) = 100$.
- In example 3, we simulate 50 data sets, each with $n = 100$ and $p = 40$ from $y = X\beta + \sigma\epsilon$ where $\epsilon \sim N(0, 1)$.

We set $\beta = (0, ..., 0, 2, ..., 2, 0, ..., 0, 2, ..., 2)$, the first 10 true β coefficients are 0, next 10 are 2, next 10 are 0, and the final 10 are 2. $\sigma = 15$ and $cond(X) = 100$.

Case $p >> n$:

In this case, as suggested by Zou (2006), we consider l_2 norm instead of OLS to compute the adaptive weights of the adaptive-lasso method. For the high-dimensional data, we simulate 30 independent data sets from $y = X\beta + \sigma\epsilon$ where $\epsilon \sim N(0, 1)$ for each of the following examples.

- In example 4, we consider $n = 200$, $p = 400$ and set all β-coefficients to be 0.85, $\sigma = 3$ and $cond(X) = 100$.
- In example 5, we consider $n = 30$ and $p = 60$, $\beta = (3, ..., 3, 0, ..., 0, 9, ..., 9)$, the first 20 true β coefficients are 3, next 20 are 0 and the next 20 are 9, $\sigma = 9$ and $cond(X) = 100$.
- In example 6, we consider $n = 100$, $p = 400$, $\beta = (1.5, 2.5)$ the first 200 true beta coefficients are 1.5 and the next 200 are 2.5, $\sigma = 9$ and $cond(X) = 100$.

Each example was considered separately. Ridge, lasso, adaptive-lasso, elastic-net and the proposed method (Huber) were fitted to the same data set simultaneously. The regularization parameter λ was set as a grid of values in the range of $[10^{-1}, 10]$. Optimal λ was determined by searching through the grid of λ values that optimize the minimization problem.

Example 1

Method	Parameters	Average MSE (*SD*)	Average of nonzero coefficients
Ridge	$\alpha = 0$	20.87 (1.90)	All
Lasso	$\alpha = 1$	20.72 (1.96)	6.74
Ad-lasso	$\gamma = 0.5$	22.18 (2.45)	6.68
	$\gamma = 1$	22.94 (2.84)	6.54
	$\gamma = 1.5$	22.96 (3.01)	6.44
Elastic-net	$\alpha = 0.25$	20.74 (1.89)	7.40
	$\alpha = 0.5$	20.62 (1.88)	7.22
	$\alpha = 0.75$	20.53 (1.87)	7.00
Huber	$\sigma = 1$	20.65 (1.88)	7.58
	$\sigma = 1.5$	20.71 (1.86)	7.58
	$\sigma = 1.8$	20.73 (1.86)	7.56

Example 2

Method	Parameters	Average MSE (*SD*)	Average of nonzero coefficients
Ridge	$\alpha = 0$	20.74 (2.62)	All
Lasso	$\alpha = 1$	21.17 (2.72)	7.28
Ad-lasso	$\gamma = 0.5$	22.56 (2.70)	7.16
	$\gamma = 1$	23.47 (2.74)	7.04
	$\gamma = 1.5$	23.82 (2.77)	6.96
Elastic-net	$\alpha = 0.25$	20.84 (2.64)	7.82
	$\alpha = 0.5$	20.91 (2.66)	7.64
	$\alpha = 0.75$	21.07 (2.69)	7.52
Huber	$\sigma = 1$	20.59 (2.61)	7.86
	$\sigma = 1.5$	20.38 (2.58)	7.94
	$\sigma = 1.8$	20.38 (2.58)	7.94

Example 3

Method	Parameters	Average MSE (*SD*)	Average of nonzero coefficients
Ridge	$\alpha = 0$	530.67 (23.29)	All
Lasso	$\alpha = 1$	563.66 (27.09)	37.48
Ad-lasso	$\gamma = 0.5$	584.03 (28.84)	37.84
	$\gamma = 1$	597.43 (29.50)	37.42
	$\gamma = 1.5$	604.28 (29.72)	37.06
Elastic-net	$\alpha = 0.25$	536.13 (23.85)	38.80
	$\alpha = 0.5$	542.77 (24.61)	38.64
	$\alpha = 0.75$	551.43 (25.61)	38.94
Huber	$\sigma = 1$	539.01 (24.35)	38.92
	$\sigma = 1.5$	524.07 (22.73)	39.02
	$\sigma = 1.8$	517.99 (22.21)	39.90

Example 4

Method	Average MSE (*SD*)	Average of nonzero coefficients
Ridge	1049.90 (38.89)	All
Lasso	1613.59 (67.23)	129.66
Ad-lasso	1627.17 (63.23)	120.30
Elastic-net	1208.14 (44.20)	253.93
Huber	1050.82 (39.03)	396.66

Example 5

Method	Average MSE (*SD*)	Average of nonzero coefficients
Ridge	39185.26 (3356.50)	All
Lasso	62093.76 (5021.50)	18.96
Ad-lasso	68142.08 (5149.02)	19.00
Elastic-net	39045.61 (3448.09)	51.16
Huber	52078.32 (3771.15)	59.80

Example 6

Method	Average MSE (*SD*)	Average of nonzero coefficients
Ridge	6015.75 (379.099)	All
Lasso	8071.26 (445.008)	65.73
Ad-lasso	8497.87 (464.066)	65.16
Elastic-net	6178.55 (398.025)	242.00
Huber	6140 (368.709)	398.00

From examples 1–3, simulations confirmed that the lasso and ad-lasso identified small number of important predictors. We observed that Huber method achieved high predictive accuracy. When the number of predictors was very large, the accuracy of Huber model was higher compared to ridge, lasso, adaptive-lasso and elastic-net. These confirmed the superiority of Huber, significantly, in the case of strong collinearity with high error amplitude. Note that the proposed method benefits from the property of sparsity in some cases, by shrinking some coefficients to zero through the l1 penalty. We observed that ridge and elastic-net generally improved over the lasso. The lasso is performing better than adaptive-lasso, however, the adaptive-lasso tends to select more variables than the lasso. Finally, in $n > p$ case, high predictive accuracy was observed with Huber method followed by ridge regression. Eventually, elastic-net, lasso and adaptive-lasso incorporate variable selection more than Huber and their resulting model was more interpretable than that of the proposed method. From examples 4–6, which represent the high-dimensional case, we observed that ridge regression and Huber outperform all the methods in example 4 and example 6 in terms of prediction accuracy. Elastic-net is performing better than other methods in example 5. Lasso and adaptive-lasso tend to select more variables than the other methods. In a high-dimensional case, we observed that there is no procedure that statistically outperforms all the others. This expected as different procedures could be used in different settings in practice.

4 | APPLICATION: PREDICTING RECOVERY AFTER STROKE

4.1 | Data and modelling approach

We used stroke data published by Douiri et al. (2017) to evaluate the proposed hybrid regularization method as well other published methods mentioned above. Data collected include 495 patients from the population-based SLSR between the period August 2002 and October 2004. All patients with cerebral infractions (ICD-10 code 163), were included. The

progression of functional recovery was evaluated using Barthel Index (BI) with total possible scores ranging from 0 to 20, with lower scores indicating increased disability. BI was measured at weeks 1, 2, 3, 4, 6, 8, 12, 26 and 52 after stroke. A number of candidate predictors were considered in the model variable selection, including demographics (age, sex, ethnicity, premorbid disability, socioeconomic status), stroke characteristics (subtype based on the Oxford classification (lacunar infarct, LACI), total anterior circulation infarcts (TACI), partial anterior circulation infarcts (PACI), posterior circulation infarcts (POCI) and intra-cerebral haemorrhagic (ICH) stroke, presence of cerebellar symptoms, baseline impairments, case-mix variables (Glasgow coma score, GCS), National Institutes of Health Stroke Scale (NIHSS)). Potential variables of recovery were screened for practicality based on their prevalence in academic literature, resulting in seven candidates prognostic factors. Details of these predictors can be found (Douiri, Grace & Sarker, 2017). Please note that all methods discussed are robust for collinearity.

In relation to missing data and based on the assumption that these were at random, we used a multiple imputation method known as Markov Chain Monte Carlo (MCMC) (Gilks, Richardson, & Spiegelhalter, 1995). The seven predictors (sex (sex), ethnic groups (ethgrp), Glasgow coma score (glas_cs), stroke subtype (subtype: TACI, PACI, LACI, POCI), Patient age (age), first week BI score (batotw1) and NIHSS Stroke Scale (nihtot) in order to predict outcomes of BI at 12, 26 and 52 weeks. Ridge, lasso, adaptive-lasso, elastic-net and huber methods were applied to the stroke data. Model fitting and tuning parameter selection that optimize the minimization problem were done on the training data. We then compared the performance of the aforementioned methods by computing their prediction mean squared error. Parameters estimation, average mean squared errors (MSE) and standard deviation (*SD*) were calculated by bootstrap methodology using 500 replications. As we mentioned earlier, the conditioning related to the covariance matrix for stroke data (SLSR) was calculated and is higher than 100 which indicates the existence of a strong collinearity.

Stroke data (12 weeks)

	Methods				
Predictor	**Ridge (sd)**	**Lasso (*SD*)**	**Ad-lasso (*SD*)**	**Elastic-net (*SD*)**	**Huber (*SD*)**
Coefficients estimations					
Sex	−1.06 (0.02)	−1.08 (0.03)	−1.23 (0.04)	−1.07 (0.03)	−0.84 (0.02)
ethgrp 1	1.18 (0.02)	1.33 (0.04)	2.16 (0.08)	1.24 (0.03)	0.86 (0.02)
ethgrp 2	0.08 (0.02)	0.17 (0.02)	1.25 (0.08)	0.13 (0.02)	−0.08 (0.02)
ethgrp 3	−0.10 (0.02)	−0.01 (0.01)	0.20 (0.11)	−0.06 (0.01)	−0.06 (0.01)
ethgrp 4	−0.24 (0.02)	−0.01 (0.00)	−0.50 (0.12)	−0.11 (0.01)	−0.15 (0.01)
glas_cs	0.61 (0.01)	0.60 (0.01)	0.51 (0.01)	0.61 (0.01)	0.61 (0.01)
TACI	−1.61 (0.02)	−1.89 (0.04)	−2.42 (0.05)	−1.72 (0.03)	−1.29 (0.02)
PACI	1.19 (0.02)	1.25 (0.03)	1.19 (0.05)	1.22 (0.02)	0.88 (0.01)
POCI	0.21 (0.04)	0.15 (0.05)	−0.28 (0.08)	0.19 (0.04)	0.18 (0.02)
LACI	−0.46 (0.03)	−0.49 (0.03)	−1.14 (0.06)	−0.46 (0.03)	−0.25 (0.02)
age	−0.04 (0.00)	−0.04 (0.00)	−0.03 (0.00)	−0.04 (0.00)	−0.04 (0.00)
batotw1	0.63 (0.00)	0.63 (0.00)	0.62 (0.00)	0.63 (0.00)	0.63 (0.00)
nihtot	0.22 (0.00)	0.21 (0.00)	0.19 (0.00)	0.22 (0.00)	0.21 (0.00)
Average MSE (*SD*)	30.66 (0.24)	31.33 (0.25)	33.46 (0.29)	30.79 (0.24)	30.34 (0.24)

Stroke data (26 weeks)

	Methods				
Predictor	**Ridge (*SD*)**	**Lasso (*SD*)**	**Ad-lasso (*SD*)**	**Elastic-net (*SD*)**	**Huber (*SD*)**
Coefficients estimations					
Sex	−0.36 (0.03)	−0.31 (0.03)	−0.18 (0.03)	−0.34 (0.03)	−0.29 (0.02)
ethgrp 1	1.08 (0.03)	1.23 (0.05)	2.58 (0.11)	1.12 (0.04)	0.79 (0.02)

(continued)

Stroke data (26 weeks) (continue)

	Methods				
Predictor	**Ridge (*SD*)**	**Lasso (*SD*)**	**Ad-lasso (*SD*)**	**Elastic-net (*SD*)**	**Huber (*SD*)**
ethgrp 2	0.13 (0.02)	0.18 (0.03)	1.60 (0.01)	0.13 (0.02)	−0.02 (0.01)
ethgrp 3	0.36 (0.01)	0.00 (0.00)	3.35 (0.15)	0.14 (0.01)	0.18 (0.01)
ethgrp 4	−0.30 (0.02)	−0.02 (0.01)	−0.29 (0.14)	−0.17 (0.01)	−0.19 (0.01)
glas_cs	0.74 (0.01)	0.72 (0.01)	0.61 (0.01)	0.73 (0.01)	0.75 (0.01)
TACI	−0.67 (0.02)	−0.48 (0.03)	−0.67 (0.04)	−0.59 (0.02)	−0.54 (0.02)
PACI	1.47 (0.02)	2.01 (0.04)	2.42 (0.05)	1.65 (0.03)	1.04 (0.01)
POCI	0.09 (0.03)	0.09 (0.04)	−0.04 (0.07)	0.09 (0.03)	0.08 (0.02)
LACI	−0.27 (0.03)	−0.19 (0.03)	−0.47 (0.05)	−0.23 (0.03)	−0.16 (0.02)
age	−0.06 (0.00)	−0.06 (0.00)	−0.05 (0.00)	−0.06 (0.00)	−0.06 (0.00)
batotw1	0.61 (0.00)	0.61 (0.00)	0.60 (0.00)	0.61 (0.00)	0.61 (0.00)
nihtot	0.18 (0.00)	0.18 (0.00)	0.14 (0.00)	0.18 (0.00)	0.18 (0.00)
Average MSE (*SD*)	31.26 (0.26)	31.53 (0.26)	33.81 (0.33)	31.30 (0.26)	30.76 (0.25)

Stroke data (52 weeks)

	Methods				
Predictor	**Ridge (*SD*)**	**Lasso (*SD*)**	**Ad-lasso (*SD*)**	**Elastic-net (*SD*)**	**Huber (*SD*)**
Coefficients estimations					
Sex	−0.52 (0.03)	−0.54 (0.03)	−0.61 (0.04)	−0.52 (0.03)	−0.39 (0.02)
ethgrp 1	1.02 (0.03)	1.08 (0.04)	1.46 (0.08)	1.03 (0.03)	0.75 (0.02)
ethgrp 2	0.17 (0.02)	0.18 (0.03)	0.87 (0.08)	0.16 (0.02)	0.01 (0.02)
ethgrp 3	0.14 (0.01)	0.00 (0.00)	0.58 (0.06)	0.00 (0.00)	0.06 (0.00)
ethgrp 4	−0.50 (0.02)	−0.04 (0.01)	−2.43 (0.12)	−0.36 (0.01)	−0.29 (0.01)
glas_cs	0.54 (0.01)	−0.52 (0.01)	0.42 (0.01)	0.54 (0.01)	0.55 (0.01)
TACI	−0.62 (0.02)	−0.37 (0.03)	−0.33 (0.03)	−0.53 (0.03)	−0.52 (0.02)
PACI	1.60 (0.02)	2.35 (0.04)	3.09 (0.05)	1.83 (0.03)	1.12 (0.02)
POCI	0.91 (0.04)	1.43 (0.07)	1.96 (0.10)	1.04 (0.05)	0.63 (0.03)
LACI	−0.51 (0.03)	−0.32 (0.04)	−0.39 (0.06)	−0.45 (0.03)	−0.37 (0.02)
age	−0.04 (0.00)	−0.04 (0.00)	−0.02 (0.00)	−0.04 (0.00)	−0.04 (0.00)
batotw1	0.68 (0.00)	0.68 (0.00)	0.67 (0.00)	0.68 (0.00)	0.68 (0.00)
nihtot	0.23 (0.00)	0.22 (0.00)	0.18 (0.00)	0.22 (0.00)	0.23 (0.00)
Average MSE (*SD*)	32.28 (0.27)	33.15 (0.29)	34.32 (0.32)	32.53 (0.27)	31.82 (0.27)

Results from these simulations confirmed that the proposed method (huber) is robust in terms of prediction accuracy. Ridge method is performing better than elastic-net, lasso and adaptive-lasso. Compared to the proposed method, the prediction accuracy of both the lasso and adaptive-lasso, is clearly affected as expected by the high conditioning number of the related covariance matrix. However, the proposed method (huber) overcomes this problem better.

5 | DISCUSSION

We have proposed an inverse problem framework to the regression problem. This approach allowed us to characterize the regularization function that optimize the regression model. A generalized penalized function was construed and based

upon robust statistics methodology. We proposed a simple but efficient hybrid regularization function. The resulting regularized regression model showed good performances compared to other known methods including ridge, lasso, adaptive-lasso and elastic-net methods. To the best of our knowledge, this is the first time where a method has used combined inverse problem approach with robust statistics methodology to solve regularized regression problem. In our simulations, we have tested the proposed method using different scenarios: error amplitudes, strong degree of collinearity which affect the conditioning of the related covariance matrix and number of predictors. The proposed method fitted well the regression model and showed good performance and less bias compared to other methods. This confirms that the inverse problem framework coupled with robust statistics methodology, has assisted to characterize the optimal regularization and was more robust compared to other methods as we observed in these simulations. Note that this proposed method could be considered as a nonlinear form of elastic-net which compromise between ridge and lasso. Using clinical data from a population-based study from London which reflect the real-world application, the method performed well in constructing regression parameters that fit well the observed data with minimal predictive error. The proposed method established an added clinical value in practice and confirms that the inverse problem framework used is optimal. In clinical research, the developed predictive model could be applied as a tool in assessing the beneficial effects of evidence-based interventions and support care settings. As a research tool, this could be used to test novel interventions or to identify enriched samples, reducing the reliance on the need for expensive and often impractical randomized controlled trials. This predictive enrichment strategy is of importance for designing future trials as it enables the enrolment of the most suitable patients thereby permitting the use of a smaller study population. Another potential application could be to derive a set of preliminary cost weights on resource uses which could help to build personalized patient care and funding models.

This approach could be extended to solve generalized mixed regression models, high dimension reduction and could also be used in statistical machine learning problems to optimize loss function (Poggio & Rosasco, 2016). The method offers all the advantages that lasso, and elastic net penalty provide for model fitting and variable selection or automated feature extraction, as well as give an understanding of the limitations related to the data. As in lasso, the proposed method also defines a continuous shrinking operation that can produce coefficients that are numerically very small (so the predictor related could be opted out). The supportive results reported here suggest that the inverse problem approach could be useful in a wide variety of statistical estimation problems. This approach is not well explored in statistical community and further study is needed to investigate further these potentials. Inverse problem framework tool could provide a useful tool for clinical and public health research. Our work added further evidence that the penalized regression using inverse problem approach is robust with less bias and could be used with assurance in the practice. Furthermore, the proposed method offers an improved alternative method to ridge, lasso, adaptive-lasso and elastic-net, which have already been shown to be valuable methods in previous studies.

ACKNOWLEDGEMENTS

AD and CW would like to acknowledge the support and funding from the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South London at King's College Hospital NHS Foundation Trust and the Royal College of Physicians, as well as the support from the NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.

YH gratefully acknowledges the financial support of the International PhD Program for Modelling Complex Systems supported by both, Sorbonne University and IRD.

CONFLICT OF INTEREST

The autors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The study and its consent procedure were approved by the ethics committees of Guy's and St Thomas' Hospital Trust, King's College Hospital, Queen's Square and Westminster Hospital. Consent for data sharing was not obtained from study participants. The research team will consider reasonable requests for sharing of anonymized patient-level data.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Youssef Hbid https://orcid.org/0000-0002-6472-7255

REFERENCES

Aitken, A. C. (1936). IV. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, *55*, 42–48.

Aldrich, J. (2005). Fisher and regression. *Statistical Science*, *20*(4), 401–417.

Baltagi (2008). *Violations of the classical assumptions*. Berlin Heidelberg: Springer.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980) *Regression diagnostics: identifying influential data and sources of collinearity*. New York: Wiley.

Bradford Hill, S. A. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300.

Broemeling, L. D. (1985). *Bayesian analysis of linear models*. New York: Marcel Dekker.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications* (1st ed.). Berlin: Springer Publishing Company, Incorporated.

Douiri, A., Grace, J., Sarker, S., Tilling, K., McKevitt, C., Wolfe, C. D., & Rudd, A. G. (2017). Patient-specific prediction of functional recovery after stroke. *International Journal of Stroke*, *12*, 539–548.

Douiri, A., Schweiger, M., Riley, J., & Arridge, S. (2005). Local diffusion regularization method for optical tomography reconstruction by using robust statistics. *Optics Letters*, ThD6. https://doi.org/10.1364/ECBO.2005.ThD6

Fu, A., Narasimhan, B., & Boyd, S. (2017). *CVXR: An R package for disciplined convex optimization*, Technical report, Department of Statistics, Stanford University.

Gauss, C. F., & Davis, C. H. (1857). *Theory of the motion of the heavenly bodies moving about the sun in conic sections a translation of Gauss's "Theoria motus."* Boston, MA: Little Brown and Company.

Geman, S. (1988). Stochastic relaxation methods for image restoration and expert systems. In G. J. Erickson & C. R. Smith (Eds.), *Maximum-entropy and Bayesian methods in science and engineering* (pp. 265–311). The Netherlands: Springer.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC Interdisciplinary Statistics, Taylor & Francis.

Hadamard, J. (1902). *Sur les problemes aux derviees partielles et leur signification physique*. Princeton, NJ: Univ Princeton Bull.

Hansen, P. C. (2001). The L-curve and its use in the numerical treatment of inverse problems. *Computational Inverse Problems in Electrocardiology*, *4*, 119–142.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, Series B*, (12), 55–70.

Huber, P. J. (1981). *Robust statistics*. New York: John Wiley and Sons.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). *Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series*. Boca Raton, FL: Chapman & Hall.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics. Themes in modern econometrics*. Cambridge, MA: Cambridge University Press.

Poggio, T., & Rosasco, L. (2016). *Course slides and videos from MIT 9.520: Statistical learning theory and applications*.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Unger, M., Pock, T., Werlberger, M., & Bischof, H. (2010). *A convex approach for variational super-resolution*. Berlin, Heidelberg: Springer.

Vogel, C. R. (2002). *Computational methods for inverse problems*. SIAM. https://epubs.siam.org/doi/abs/10.1137/1.9780898717570

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *67*(2), 301–320.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Hbid Y, Mohamed K, Wolfe CDA, Douiri A. Inverse problem approach to regularized regression models with application to predicting recovery after stroke. *Biometrical Journal*. 2020;1–13. https://doi.org/10.1002/bimj.201900283

APPENDIX

Proof A.1. Suppose that all the Hypothesis are fulfilled, (iv) affirm that $\int_\Omega \psi(|\beta(t)|)\,\mathrm{d}t$ is convex and that ensure the convexity of the functional $E(\beta)$, hence the the problem have a global minimum solution.

Let us compute the derivative of $E(\beta)$:

To deal with the non differentiability of the absolute value at t = 0, we consider: $|\beta| = \sqrt{|\beta|^2 + \epsilon}$

We have: $D\left(\beta \longrightarrow \int_\Omega \frac{1}{2}|X\beta(t) - Y|^2\,\mathrm{d}t\right) = \int_\Omega [X^t(X\beta - Y)]$

and: $D(\beta \longrightarrow \int_\Omega \psi(|\beta|)\mathrm{d}t) = \frac{d}{dt}[\int \psi(|\beta + th|)]_{t=0}$

We set: $\eta(t) = |\beta + th|^2 = |\beta|^2 + 2t\beta h + t^2.|h|^2$

and $\eta'(t) = 2\beta h + 2t|h|^2$

then $\frac{d}{dt}[\int \psi(\sqrt{\eta(t)})]_{t=0} = \frac{1}{2}\int \psi'(\sqrt{\eta(0)})\frac{\eta'(0)}{\sqrt{\eta(0)}} = \int \psi'(|\beta|)\frac{\beta}{|\beta|}h$

Hence, $\int_\Omega [X^t(X\beta - Y)] + \lambda\psi'(|\beta|)\frac{\beta}{|\beta|}h = 0$

We conclude that: $E'(\beta) = X^t(X\beta - Y) + \lambda\frac{\psi'(|\beta|)}{|\beta|}\beta$

from which the proposition follows.