Regularizing Large Biosignals with Finite Differences

Georgios Drakopoulos[1] and Vasileios Megalooikonomou[1]

Multidimensional Data Analysis and Knowledge Management (MDAKM) Lab
Computer Engineering and Informatics Department (CEID)
University of Patras, Achaia 26504, Hellas
{drakop, vasilis}@ceid.upatras.gr

Abstract. In the biomedical analytics pipeline data preprocessing is the first and crucial step as subsequent results and visualization depend heavily on original data quality. However, the latter often contain a large number of outliers or missing values. Moreover, they may be corrupted by noise of unknown characteristics. This is in many cases aggravated by lack of sufficient information to construct a data cleaning mechanism. Regularization techniques remove erroneous values and complete missing ones while requiring little or no information regarding either data or noise dynamics. This paper examines the theory and practice of a regularization class based on finite differences and implemented through the conjugate gradient method. Moreover, it explores the connection of finite differences to the discrete Laplace operator. The results obtained from applying the proposed regularization techniques to heart rate time series from the MIT-BIH dataset are discussed.

Keywords: Finite difference matrix, regularization, biosignal processing, big data analytics, conjugate gradient, discrete Laplace operator, electrocardiogram, heartbeat rate

1 Introduction

As big data displace traditional storage and processing bounds, the need for high quality data emerges as a pressing issue. Particularly in deep learning, bioengineering, and big data analytics arises frequently the problem of replacing a raw data vector **b** with a smoother or cleaner version **s**, but there is insufficient information to do so. This typically happens when **b** has been corrupted by noise of unknown distribution, possibly colored or non-stationary, excluding thus signal estimation techniques based on stationarity and gaussianity assumptions. Alternatively, **b** may contain a prohibitively large number of missing values or outliers, for instance when the original data come from databases which are either schemaless or have no integrity constraints placed on them.

A broad class of regularization techniques attempts to minimize the cost function

$$J(\mathbf{s};\gamma_0) \stackrel{\triangle}{=} \|\mathbf{b}-\mathbf{s}\|_2^2 + \gamma_0\|\mathbf{M}\mathbf{s}\|_2^2 \tag{1}$$

$\mathbf{M}$ is a constraint matrix which may well codify restrictions inherent to the data generating process, but in the majority of cases it is formed by supplemental constraints selected by a data scientist. Thus, $\mathbf{M}$ is open to interpretation. $\gamma_0 > 0$ represents the relative importance of the first term of (1), which measures how close $\mathbf{s}$ is to $\mathbf{b}$, compared to the second one, which quantifies the degree $\mathbf{b}$ conforms to the given constraints.

Once J is formed, $\mathbf{s}$ is selected to be

$$\mathbf{s}^* \;=\; \operatorname{argmin}_{\mathbf{s}} \left\{ J(\mathbf{s};\gamma_0) \right\} \tag{2}$$

A special case of (1) is when $\mathbf{M}$ is $\boldsymbol{\Delta}_p$, the p-th order finite difference matrix. In this case $J(\mathbf{s};\gamma_0)$ becomes

$$J_p(\mathbf{s};\gamma_0) \;=\; \|\mathbf{b}-\mathbf{s}\|_2^2 + \gamma_0 \|\boldsymbol{\Delta}_p \mathbf{s}\|_2^2 \tag{3}$$

When $\boldsymbol{\Delta}_p$ is applied to $\mathbf{s}$, it returns the p-th order discrete difference of $\mathbf{s}$ denoted by $\mathbf{s}_p = \boldsymbol{\Delta}_p\mathbf{s}$. When $p = 0$, then $\mathbf{s} = \mathbf{s}_0$ and $\boldsymbol{\Delta}_0$ is the identity matrix. For the cases when $p = 1$ and $p = 2$ the multiplication $\boldsymbol{\Delta}_p\mathbf{s}$ yields respectively

$$\mathbf{s}_1 \;=\; \boldsymbol{\Delta}_1\mathbf{s} \;=\; \begin{bmatrix} \mathbf{s}[1]-\mathbf{s}[0] \\ \mathbf{s}[2]-\mathbf{s}[1] \\ \vdots \\ \mathbf{s}[n-1]-\mathbf{s}[n-2] \end{bmatrix} \tag{4}$$

$$\mathbf{s}_2 \;=\; \boldsymbol{\Delta}_2\mathbf{s} \;=\; \begin{bmatrix} \mathbf{s}[2]-2\mathbf{s}[1]+\mathbf{s}[0] \\ \mathbf{s}[3]-2\mathbf{s}[2]+\mathbf{s}[1] \\ \vdots \\ \mathbf{s}[n-1]-2\mathbf{s}[n-2]+\mathbf{s}[n-3] \end{bmatrix} \tag{5}$$

The primary contribution of this paper is two families of algorithms for time series regularization based on finite differences. They require only two parameters and they can implemented efficiently as variations of the conjugate gradient method, a widespread iterative algorithm for solving symmetric and positive definite linear systems. Selected members of both families are applied to denoising two electrocardiogram (ECG) time series from the MIT-BIH dataset.

The remaining of this work is structured as follows. Section 2 summarizes current scientific literature regarding this topic. Section 3 outlines the properties of difference matrices and section 5 a special case approximation based on discrete Laplace operators. Finally, section 6 presents the results of applying the proposed regularization algorithms to cleaning brain data, whereas section 7 discusses future research directions. Table 1 summarizes the symbols used in this paper. Vectors and sequences with n elements are indexed from 0 to $n-1$. Vectors are symbolized by small boldface letters, matrices by capital boldface letters, and scalars by small Greek letters. Throughout this paper is assumed that difference order p is considerably smaller than the vector length n. Acronyms are defined the first time they are encountered in the text.

Table 1. Symbols used in this paper.

Symbol	**Meaning**
$\triangleq$	Definition or equality by definition
$\mathbf{x}^{[k]}$	k-th vector of an iterative algorithm
$\lambda(\mathbf{A})$	Spectrum of matrix $\mathbf{A}$
$q_{\mathbf{A}}(\lambda)$	Characteristic polynomial of matrix $\mathbf{A}$
$\langle x_k \rangle$	Sequence of elements x_k
$\langle x_k \rangle \star \langle y_k \rangle$	Linear convolution of $\langle x_k \rangle$ with $\langle y_k \rangle$
$\mathcal{F}\left[\mathbf{x}\right]$	Discrete Fourier transform of vector $\mathbf{x}$
$\mathcal{F}^{-1}\left[\mathbf{x}\right]$	Inverse discrete Fourier transform of $\mathbf{x}$
$\mathbf{M}_n$	Discrete Laplace operator $n \times n$
$\mathbf{x}_1 \oslash \mathbf{x}_2$	Elementwise vector division of $\mathbf{x}_1$ to $\mathbf{x}_2$
$\Re\left[\mathbf{x}\right]$	Real part of vector $\mathbf{x}$

2 Related Work

Regularization has been originally proposed for obtaining a solution from ill-posed linear systems and inverese problems as explained in the overview [15]. Among the earliest and most known techniques is Tikhonov regularization [15][26]

$$L(\mathbf{s}; \mathbf{A}, \mathbf{B}) \; = \; \left\| \mathbf{b} - \mathbf{A}\mathbf{s} \right\|_M^2 + \left\| \mathbf{B}\mathbf{s} \right\|_M^2$$

where $\left\| \cdot \right\|_M$ is a suitably selected norm, not necessarily the Euclidean, defined on a metric space. A scheme for selecting parameters for Tikhonov regularization is proposed in [22], while extensions can be found in [11] and [6]. Lasso [25] is another methodology moving along similar lines using the Chebyshev norm. Regularization properties under the Euclidean and the Chebyshev norm are discussed in [21].

An established technique for solving continous inverse where a bounded function $f(\cdot)$ is satisfying the n constraints $f(x_k) = y_k$ with $x_1 \leq x_2 \leq \ldots \leq x_n$ relies on a similar cost function L_S defined over a Sobolev space [10]

$$L_S(f; \eta_0) \; = \; \sum_{k=1}^{n} \left| f(x_k) - y_k \right|^2 + \eta_0 \int_{x_1}^{x_n} \left(\frac{\partial^p f(x)}{\partial x^p} \right)^2 dx$$

In ordinary machine learning regularization has been proposed for pruning synapses in neural networks once the training phase is complete [9]. Alternatively, in supervised deep learning regularizing neural networks [13][17][2] is an architecture designed to prevent overfitting during the training phase [24][5]. In unsupervised deep learning variants of the non-negative matrix factorization (NMF) [18] have been proposed to avoid near singular factors. Regularization algorithms such as [20] have been introduced for reproducible kernel Hilbert spaces (RKHSs) [12].

Cardiovascular time series and electrocardiograms (ECG) have long been the subject of biomedical research [14] and they are common biosignals along with respiration rate, blood samples, and sleep data. Machine learning methods have been applied to ECG data in [27], while in [19] heart rate variability was the primary feature in a broad range of predictors. For a signal processing in heart time series see [3][1]. Finally, [4] explains how ECGs can be described in terms of the Minimum Description Length (MDL) principle.

3 Difference Matrix

Definition 1. *The p-th order difference of* $\mathbf{y} \in \mathbb{R}^n$*, denoted by* $\mathbf{y}_p \in \mathbb{R}^{n-p}$*, is elementwise recursively computed as*

$$\mathbf{y}_p[k] \triangleq \begin{cases} \mathbf{y}[k], & p = 0 \\ \mathbf{y}_{p-1}[k+1] - \mathbf{y}_{p-1}[k], & p \geq 1 \end{cases} \tag{6}$$

Property 1. Definition 1 implies that p-th order difference is a linear combination of $p+1$ consecutive elements of $\mathbf{s}$. Let $\langle \tau_p[k] \rangle$ denote the corresponding coefficient sequence. Then

$$\langle \tau_p[k] \rangle = (-1)^k \binom{p}{k}, \quad 0 \leq k \leq p \tag{7}$$

Proof. Let $T_p(z)$ be the generating function of $\langle \tau[k] \rangle$. Clearly, $T_0(z) = 1$ and $T_1(z) = 1 - z$. As the p-th order difference is computed by subtracting from the $(p-1)$-th order sequence a delayed by one copy of itself. Therefore

$$\begin{aligned} T_p(z) &= T_{p-1}(z) - z\, T_{p-1}(z) \\ &= (1-z)\, T_{p-1}(z) = (1-z)^p \end{aligned} \tag{8}$$

implying that $\tau_p[k]$ is the modified binomial coefficient of (7).

Property 2. $\langle \tau_p[k] \rangle$ has the following properties:

$$\begin{aligned} \sum_{k=0}^{p} \tau_p[k] &= \sum_{k=0}^{p} (-1)^k \binom{p}{k} = 0 \\ \sum_{k=0}^{p} |\tau_p[k]| &= \sum_{k=0}^{p} \binom{p}{k} = 2^p \\ \sum_{k=0}^{p} |\tau_p[k]|^2 &= \sum_{k=0}^{p} \binom{p}{k} \binom{p}{p-k} = \binom{2p}{p} \\ \langle \tau_{p_1}[k] \rangle \star \langle \tau_{p_2}[k] \rangle &= \langle \tau_{p_1+p_2}[k] \rangle \end{aligned} \tag{9}$$

The normalized Fourier transform of $\langle \tau_p[k] \rangle$ is

$$\begin{aligned} (\mathcal{F}[\tau_p])[u] &\triangleq \frac{1}{\sqrt{p+1}} \sum_{k=0}^{p} (-1)^k \binom{p}{k} e^{-iku\frac{2\pi}{p+1}} \\ &= \frac{2^p \sin^p \left(\frac{u\pi}{p+1}\right)}{\sqrt{p+1}} e^{ip\pi\left(\frac{1}{2} - \frac{u}{p+1}\right)} \end{aligned} \tag{10}$$

Property 3. $\langle \tau_p[k] \rangle$ is a linear phase FIR filter.

$\mathbf{\Delta}_p$ can be directly constructed from $\langle \tau_p[k] \rangle$ as follows

$$\underbrace{\begin{bmatrix} \tau_p[p] & \tau_p[p-1] & \ldots & \tau_p[0] & 0 & \ldots & 0 \\ 0 & \tau_p[p] & \ldots & \tau_p[1] & \tau_p[0] & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ldots & \vdots \\ 0 & \ldots & \ldots & \ldots & \tau_p[p] & \ldots & \tau_p[0] \end{bmatrix}}_{\triangleq \mathbf{\Delta}_p} \tag{11}$$

Property 4. By construction $\mathbf{\Delta}_p \in \mathbb{R}^{(n-p)\times n}$. Moreover

- It is non-normal as $\mathbf{\Delta}_p^T \mathbf{\Delta}_p \neq \mathbf{\Delta}_p \mathbf{\Delta}_p^T$.
- It is a filter matrix.
- It is Toeplitz and upper triangular with a band of p.
- It has $(n-p)(p+1) = \mathrm{O}\,(n)$ non-zero elements.
- $\|\mathbf{\Delta}_p\|_1 = \|\mathbf{\Delta}_p\|_\infty = \sum_{k=0}^{n} |\tau_p[k]| = 2^p$
- $\|\mathbf{\Delta}_p\|_F = \sqrt{(n-p)\sum_{k=0}^{p} |\tau_p[k]|^2} = \sqrt{(n-p)\binom{2p}{p}}$

Property 5. By construction $\mathbf{\Delta}_p^T \mathbf{\Delta}_p \in \mathbb{R}^{n\times n}$. Moreover

- It is p-diagonal. Also is symmetric and, hence, normal.
- It is positive semidefinite with a nullity of p.
- $\left\|\mathbf{\Delta}_p^T \mathbf{\Delta}_p\right\|_1 = \left\|\mathbf{\Delta}_p^T \mathbf{\Delta}_p\right\|_2 = \sum_{k=0}^{p} \tau_p^2[k] = \binom{2p}{p}$
- $(n-2p)\binom{2p}{p} \leq \left\|\mathbf{\Delta}_p^T \mathbf{\Delta}_p\right\|_F \leq (n-p)\binom{2p}{p}$

4 Finite Difference Regularization

As $J_p(\mathbf{s};\gamma_0)$ is quadratic in $\mathbf{s}$, its minimum can be found through its Jacobian $\nabla_{\mathbf{s}} J_p(\mathbf{s};\gamma_0)$ and Hessian $\nabla^2_{\mathbf{s}} J_p(\mathbf{s};\gamma_0)$.

$$\begin{aligned} J_p(\mathbf{s};\gamma_0) &= \|\mathbf{b}-\mathbf{s}\|_2^2 + \gamma_0 \|\mathbf{\Delta}_p \mathbf{s}\|_2^2 \\ &= \mathbf{b}^T\mathbf{b} - 2\mathbf{s}^T\mathbf{b} + \mathbf{s}^T\left(\mathbf{I}_n + \gamma_0 \mathbf{\Delta}_p^T \mathbf{\Delta}_p\right)\mathbf{s} \\ \nabla_{\mathbf{s}} J_p(\mathbf{s};\gamma_0) &= 2\big(\mathbf{I}_n + \gamma_0 \mathbf{\Delta}_p^T \mathbf{\Delta}_p\big)\mathbf{s} - 2\mathbf{b} \\ \nabla^2_{\mathbf{s}} J_p(\mathbf{s};\gamma_0) &= \mathbf{I}_n + \gamma_0 \mathbf{\Delta}_p^T \mathbf{\Delta}_p \triangleq \mathbf{K}_p \end{aligned} \tag{12}$$

When $J_p(\mathbf{s};\gamma_0)$ is constant, its contours are hyperellipses whose foci are $\mathbf{b}$ and $\mathbf{0}$.

Property 6. $\mathbf{K}_p \in \mathbb{R}^{n\times n}$ has the following properties:

- It is symmetric and, therefore, normal.
- It is p-diagonal.

– It is positive definite as for any non-zero vector $\mathbf{u}$

$$\mathbf{u}^T\left(\mathbf{I}_n + \gamma_0 \boldsymbol{\Delta}_p^T \boldsymbol{\Delta}_p\right)\mathbf{u} \;=\; \|\mathbf{u}\|_2^2 + \gamma_0 \|\boldsymbol{\Delta}_p \mathbf{u}\|_2^2 \;>\; 0$$

Since $\mathbf{K}_p$ is positive defnite, $J_p(\mathbf{s};\gamma_0)$ has a unique global minimum which can be found by zeroing $\nabla_{\mathbf{s}} J_p(\mathbf{s};\gamma_0)$. Then

$$\mathbf{s}_p^* \;=\; \mathbf{K}_p^{-1}\mathbf{b} \;=\; \left(\mathbf{I}_n + \gamma_0 \boldsymbol{\Delta}_p^T \boldsymbol{\Delta}_p\right)^{-1}\mathbf{b} \tag{13}$$

Since (13) is large and sparse, an iterative solver is preferable to a direct one in terms of both computational cost and memory. $\mathbf{K}_p$ is symmetric, positive definite, and sparse by construction and, therefore, the conjugate gradient algorithm can be used. Its general form is analyzed in detail in [23][7] for clarity, whereas its special from for the particular regularization problem is outlined in algorithm 1.

Since $\mathbf{K}_p$ has n eigenvalues, then at most n iterations are required [23]. However, if the algebraic multiplicity of some eigenvalues is larger than one or when eigenvalues tend to be clustered, then less iterations are needed [7]. Observe that $\mathbf{K}_p$ need not and should not be explicitly constructed as $\mathbf{K}_p\mathbf{s}$ can be computed in $\mathrm{O}\left(n\right)$ time knowing only $\mathbf{s}$ and $\langle \tau_p[k] \rangle$. Therefore, only few iterations may be required and they are computationally efficient.

Algorithm 1 Conjugate Gradient with Finite Differences

```
Require: Data $\mathbf{b}$, guess $\mathbf{s}_p^{[0]}$, threshold $\eta_0 > 0$, integer $p > 0$
Ensure: $\mathbf{s}_p^{[k+1]}$ is an approximate solution of $\mathbf{K}_p\mathbf{s} \;=\; \mathbf{b}$
compute (or retrieve from a lookup table) $\langle \tau_p[k] \rangle$
$\mathbf{r}^{[0]} \;\leftarrow\; \mathbf{b} - \left(\mathbf{s}_p^{[0]} + \gamma_0 \boldsymbol{\Delta}_p^T \left(\boldsymbol{\Delta}_{\mathbf{p}} \mathbf{s}_p^{[0]}\right)\right)$ and $\mathbf{p}^{[0]} \;\leftarrow\; \mathbf{r}^{[0]}$
while $\left\|\mathbf{r}^{[k]}\right\|_2 > \eta_0$ do
    $\mathbf{g}^{[k]} \;\leftarrow\; \boldsymbol{\Delta}_p \mathbf{p}^{[k]}$
    $\alpha^{[k]} \;\leftarrow\; \left\|\mathbf{r}^{[k]}\right\|_2^2 / \left(\left\|\mathbf{p}^{[k]}\right\|_2^2 + \gamma_0 \left\|\mathbf{g}^{[k]}\right\|_2^2\right)$
    $\mathbf{s}_p^{[k+1]} \;\leftarrow\; \mathbf{s}_p^{[k]} + \alpha^{[k]} \mathbf{p}^{[k]}$
    $\mathbf{r}^{[k+1]} \;\leftarrow\; \mathbf{r}^{[k]} - \alpha^{[k]} \left(\mathbf{g}^{[k]} + \gamma_0 \boldsymbol{\Delta}_p^T \mathbf{g}^{[k]}\right)$
    $\beta^{[k]} \;\leftarrow\; \left\|\mathbf{r}^{[k+1]}\right\|_2 / \left\|\mathbf{r}^{[k]}\right\|_2$
    $\mathbf{p}^{[k+1]} \;\leftarrow\; \mathbf{r}^{[k]} + \beta^{[k]} \mathbf{p}^{[k]}$
    $k \;\leftarrow\; k + 1$
end while
return $\mathbf{s}_p^{[k+1]}$
```

Since the entries of $\mathbf{b}$ are unlabeled, the proposed regularization techniques can also be viewed as unsupervised learning algorithms. Alternatively, from a signal processing perspective, regularization can be regarded as a lowpass filter.

Also, notice that any direct solver for system (13) is another way to conduct offline data analysis, while iterative solutions are suitable for online analysis.

When p is zero, the cost function is

$$J_0(\mathbf{s};\gamma_0) \;=\; \|\mathbf{b}-\mathbf{s}\|_2^2 + \gamma_0\|\mathbf{s}\|_2^2 \tag{14}$$

indicating that a tradeoff between the approximation $\mathbf{s}$ to $\mathbf{b}$ and the overall energy of $\mathbf{s}$ is sought. This yields a smoother signal with shorter spikes. As spikes often are either a noisy outburst or outliers, this kind of smoothing is desirable.

Expanding the norms and zeroing $\nabla_{\mathbf{s}} J_0(\mathbf{s};\gamma_0)$ yields

$$\mathbf{s}_0^* \;=\; \left(\frac{1}{1+\gamma_0}\right)\mathbf{b} \tag{15}$$

Notice that for small values of γ_0 the solution $\mathbf{s}_0^*$ will be close to $\mathbf{b}$, whereas for large values will be essentially zero.

In this case the cost function is

$$J_1(\mathbf{s};\gamma_0) \;=\; \|\mathbf{b}-\mathbf{s}\|_2^2 + \gamma_0\|\boldsymbol{\Delta}_1\mathbf{s}\|_2^2 \tag{16}$$

and approximations to $\mathbf{b}$ with smooth first discrete derivative are sought. Thus, for large values of γ_0, the solution $\mathbf{s}_1^*$ will be the straight line that best fits $\mathbf{b}$ in the least squares sense.

Property 7. $\lambda(\mathbf{K}_1)$ remains bounded with n and

$$1 \;\leq\; \lambda(\mathbf{K}_1) \;\leq\; 1+4\gamma_0 \tag{17}$$

Proof. By application of the Gershgorin theorem.

Property 8. $q_{\mathbf{K}_1}(\lambda)$ can be shown to be

$$\begin{aligned} q_{\mathbf{K}_1}(\lambda) \;=\; & (1-\lambda)\,\arccos\left((n-1)\cos\left(\frac{2-\lambda}{2}\right)\right) \\ & -\arccos\left((n-2)\cos\left(\frac{2-\lambda}{2}\right)\right) \end{aligned} \tag{18}$$

Proof. By induction on the determinant of $\mathbf{K}_1 - \lambda\mathbf{I}_n$.

The cost function when $p=2$ becomes

$$J_2(\mathbf{s};\gamma_0) \;=\; \|\mathbf{b}-\mathbf{s}\|_2^2 + \gamma_0\|\boldsymbol{\Delta}_2\mathbf{s}\|_2^2 \tag{19}$$

In this case, approximations to the data set $\mathbf{s}$ are sought so that their second discrete derivative is as smooth as possible.

Property 9. $\lambda(\mathbf{K}_2)$ remains bounded with n and

$$1-4\gamma_0 \;\leq\; \lambda(\mathbf{K}_2) \;\leq\; 1+16\gamma_0 \tag{20}$$

Proof. By application of the Gershgorin theorem.

Property 10. $q_{\mathbf{K}_2}(\lambda)$ is upper bounded by the Chebyshev polynomial of first kind of order n defined as

$$T_n(\lambda) \triangleq \begin{cases} \cos\left(n \arccos\left(\lambda\right)\right), & |\lambda| \leq 1 \\ \cosh\left(n \operatorname{arccosh}\left(\lambda\right)\right), & \lambda > 1 \\ \left(-1\right)^n \cosh\left(n \operatorname{arccosh}\left(-\lambda\right)\right), & \lambda < -1 \end{cases} \tag{21}$$

where $\left\| q_{\mathbf{K}_2}(\lambda) - T_n(\lambda) \right\|_\infty \to 0$ as $n \to +\infty$.

Proof. By induction on the determinant of $\mathbf{K}_2 - \lambda \mathbf{I}_n$ and observing the dominant terms of both polynomials.

Cost function (12) relies on minimal knowledge regarding $\mathbf{b}$, coded implicitly through p and γ_0. Choosing p is equivalent to model selection. Although methodologies such as the MDL are available, two techniques relevant to the specific problem are examined. One way of selecting p lies in bounding the spectral content of $\boldsymbol{\Delta}_p \mathbf{b}$. This matrix-vector multiplication is tantamount to differentiating $\mathbf{b}$ p times and, therefore, increasing the power in high frequencies proportionally to that frequency. Another way is to simply bound the first derivative, as the heartbeat rate is primarily a periodic signal. Selecting γ_0 is more straightforward as it is bounded by the spectral requirements of either coefficient matrices. Also, γ_0 can be chosen as the inverse of a fraction of the data vector power.

5 Laplace Operator Approximation

The structure of $\boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1$ is very similar to that of $-\mathbf{M}_n$, where $\mathbf{M}_n$ denotes the $n \times n$ discrete Laplace operator defined as

$$\mathbf{M}_n \triangleq \operatorname{trid}\begin{bmatrix} -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{n \times n} \tag{22}$$

Their nonzero patterns were identical as both are tridiagonal matrices. Moreover, for $16 \leq n \leq 8192$

$$\left\| \mathbf{M}_n + \boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1 \right\|_1 = \left\| \mathbf{M}_n + \boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1 \right\|_2 = 1 \tag{23}$$

This is attributed to the fact that their difference has only two entries of value 1 regardless of n. Besides norms, another way to examine matrix similarity is the behavior of their spectra. Let $\{\lambda_k\}$ and $\{\lambda'_k\}$ denote the sets of sorted eigenvalues of $-\mathbf{M}_n$ and $\boldsymbol{\Delta}_1^T \boldsymbol{\Delta}_1$ respectively. Also let $\{\mathbf{g}_k\}$ and $\{\mathbf{g}'_k\}$ denote the sets of corresponding eigenvectors. Then

$$\nu^* \triangleq \frac{1}{n} \sqrt{\sum_{k=1}^{n} \left(\lambda_k - \lambda'_k\right)^2} \text{ and } \mu^* \triangleq \frac{1}{n} \sum_{k=1}^{n} \left\| \mathbf{g}_k - \mathbf{g}'_k \right\|_2 \tag{24}$$

Table 2. Matrix approximation metrics.

n	ν^*	μ^*	n	ν^*	μ^*
16	0.0363	0.3558	512	0.0002	0.0465
32	0.0129	0.1739	1024	0.00007	0.0333
64	0.0046	0.1348	2048	0.00002	0.0235
128	0.0016	0.0937	4096	0.00001	0.0166
256	0.0006	0.0661	8192	0.000003	0.0117

are matrix approximation metrics whose values are shown in table 2. Their values, along with the clustered nature of both $\lambda(-\mathbf{M}_n)$ and $\lambda\big(\boldsymbol{\Delta}_1^T\boldsymbol{\Delta}_1\big)$, explain why ν^* and μ^* vanish with n, implying this approximation is suitable for large datasets.

Thus, $\mathbf{s}_1^*$ derived from $J_1(\mathbf{s};\gamma_0)$ can be approximated by

$$\mathbf{s}_1^* \;=\; \big(\mathbf{I}_n + \gamma_0\boldsymbol{\Delta}_1^T\boldsymbol{\Delta}_1\big)^{-1}\mathbf{b} \;\approx\; \left(\mathbf{I}_n - \gamma_0\mathbf{M}_n\right)^{-1}\mathbf{b} \tag{25}$$

Additionally, equation (25) can be alternatively expressed in terms of the Neumann matrix power series as

$$\mathbf{s}_1^* \approx \left(\mathbf{I}_n - \gamma_0\mathbf{M}_n\right)^{-1}\mathbf{b} = \sum_{k=0}^{+\infty}\left(\gamma_0\mathbf{M}_n\right)^k\mathbf{b} = \sum_{k=0}^{+\infty}\left(\gamma_0^k\mathbf{M}_n^k\mathbf{b}\right) \tag{26}$$

provided that γ_0 is selected so that the *spectral radius* of $(\mathbf{I}_n - \gamma_0\mathbf{M}_n)$, namely the maximum absolute value of its eigenvalues, lies in $(0, 1]$ [7][23]. The rightmost formulation of (26) suggests that a recursive computation thereof is feasible. In that case, care should be taken to avoid numerical errors with techniques such as those for computing vertex centrality in large graphs through adjacency matrix power series [8].

Notice that for the conjugate gradient to work $(\mathbf{I}_n - \gamma_0)$ must be positive definite. This is accomplished by selecting γ_0 such that its spectrum is strictly positive, though it is possible in certain cases that conjugate gradient works provided that $\mathbf{s}^{[0]}$ is defined over the subspace spanned by the eigenvectors corresponding only to positive eigenvalues [23].

A well known fact regarding $\mathbf{M}_n$ is that its eigenvalues are

$$\lambda_k \;=\; 2\left(1 - \cos\left(\frac{k\pi}{n+1}\right)\right), \quad 1 \,\leq\, k \,\leq\, n \tag{27}$$

and its eigenvectors form the discrete sine transform basis [7]

$$\mathbf{g}_k \;=\; \sqrt{\frac{2}{n+1}}\left[\sin\left(\tfrac{k\pi}{n+1}\right)\,\ldots\,\sin\left(\tfrac{kn\pi}{n+1}\right)\right]^T \tag{28}$$

which is the real part of the Fourier transform. As computationally efficient algorithms exist for (28), while (27) is trivially computed, it becomes clear that

Algorithm 2 Laplace Regularization (when $p = 1$)

```
Require: Data $\mathbf{b}$, guess $\mathbf{s}_1^{[0]}$, flag $T$, threshold $\eta_0 > 0$
Ensure: $\mathbf{s}_1^{[k+1]}$ is an approximate solution of $\mathbf{K}_1\mathbf{s} = \mathbf{b}$
if $T$ is true then
    $\mathbf{s}_1^{[0]} \leftarrow \mathbf{b}$ and $k \leftarrow 0$
    repeat
        $\mathbf{s}_1^{[k+1]} \leftarrow (\mathbf{I}_n + \gamma_0\mathbf{M}_n)\mathbf{s}_1^{[k]}$ and $k \leftarrow k + 1$
    until $\left\|\mathbf{s}_1^{[k]} - \mathbf{s}_1^{[k-1]}\right\|_2 > \eta_0$
else
    $\mathbf{r}^{[0]} \leftarrow \mathbf{b} - (\mathbf{I}_n - \gamma_0\mathbf{M}_n)\mathbf{s}_1^{[0]}$ and $\mathbf{p}^{[0]} \leftarrow \mathbf{r}^{[0]}$
    while $\|\mathbf{r}^{[k]}\|_2 > \eta_0$ do
        $\alpha^{[k]} \leftarrow \|\mathbf{r}^{[k]}\|_2^2 / \left(\|\mathbf{p}^{[k]}\|_2^2 - \gamma_0(\mathbf{p}^{[k]})^T\mathbf{M}_n\mathbf{p}^{[k]}\right)$
        $\mathbf{s}_1^{[k+1]} \leftarrow \mathbf{s}_1^{[k]} + \alpha^{[k]}\mathbf{p}^{[k]}$
        $\mathbf{r}^{[k+1]} \leftarrow \mathbf{r}^{[k]} - \alpha^{[k]}(\mathbf{I}_n - \gamma_0\mathbf{M}_n)\mathbf{p}^{[k]}$
        $\beta^{[k]} \leftarrow \|\mathbf{r}^{[k+1]}\|_2 / \|\mathbf{r}^{[k]}\|_2$
        $\mathbf{p}^{[k+1]} \leftarrow \mathbf{r}^{[k]} + \beta^{[k]}\mathbf{p}^{[k]}$
        $k \leftarrow k + 1$
    end while
end if
return $\mathbf{s}_1^{[k+1]}$
```

yet another an alternative for obtaining $\mathbf{s}_1^*$ based on spectral properties of $\mathbf{M}_n$ exists. To summarize, there are three main methodologies based on the discrete Laplace operator, namely the tailored conjugate gradient variant of algorithm 1, the Neuman power series expansion, and the spectral method. The first two are outlined in algorithm 2 and the last one, being of different nature, is shown separately in algorithm 3.

Algorithm 3 Spectral Laplace Regularization (when $p = 1$)

```
Require: Data vector $\mathbf{b}$
Ensure: $\mathbf{s}_1$ is an approximate solution of $\mathbf{K}_1\mathbf{s} = \mathbf{b}$
return $\Re\left[\mathcal{F}^{-1}\left[\mathcal{F}\left[\mathbf{b}\right] \oslash \left[1 - \gamma_0\lambda_1 \ldots 1 - \gamma_0\lambda_n\right]^T\right]\right]$
```

Notice that algorithm 3 despite its memory efficiency and numerical stability corresponds to a direct solver. This contrasts the iterative algorithms 1 and 2.

6 Results

A number of standard ECG datasets are available online, contaning heartbeat rates from subjects with a broad range of potentially severe heart problems such as arrythmia, tachyarrythmia, braduarrythmia and irregular heartbeat rate. These conditions usually indicate heart problems such as cardiomyopathy or perturbed control from the sinoatrial node.

Datasets include the Arrythmia dataset from the UCI online repository which is suitable for designing and testing classifiers against healthy and arrythmiac heartbeat rates based on 279 attributes [16]. The St.Petersburg Arrythmia Database consists of 75 annotated recordings extracted from 32 patients undergoing tests for coronary artery disease. Finally, the standard ECG benchmark dataset is the MIT-BIH [14] has been used. Its simplest form contains the instantaneous heart rates of two subjects engaged in comparable activities. Each series consists of 1800 evenly-spaced measurements taken every 0.5 seconds for a total of 15 minutes of activity with nearly identical mean values and standard deviations.

Noise removal has been measured by adding white Gaussian noise to the original waveforms, and then the mean squared error (MSE) between the regularized and the original versions has been computed for the two MIT-BIH time series. $\phi = \left(1+\sqrt{5}\right)/2$. The SNR was 10dB.

Table 3. MSE for additive white Gaussian noise.

γ_0	$p=0$	$p=1$	$p=1(N)$	$p=1(S)$	$p=2$
$\phi/2$	0.8119	0.6635	0.7001	0.5991	0.7719
$\phi/4$	1.3833	1.2302	1.2449	1.2395	1.6821
$\phi/2$	0.8350	0.6808	0.7794	0.7021	0.9112
$\phi/4$	1.4176	1.2525	1.3006	1.2890	1.4467

The corresponding execution times are shown in table 4.

Table 4. Time for additive white Gaussian noise (sec).

γ_0	$p=0$	$p=1$	$p=1(N)$	$p=1(S)$	$p=2$
$\phi/2$	1.5029	1.5033	3.2211	0.4017	1.5025
$\phi/4$	1.5028	1.5001	4.1144	0.4022	1.5000
$\phi/2$	1.5004	6.5013	3.9954	0.4000	1.5011
$\phi/4$	1.5009	1.5090	4.0322	0.4023	1.5009

Outlier smoothing has been performed as follows. 60 random samples have been replaced by very large values of the same sign as the replaced one and the MSE between the regularized and the original time series is computed.

Table 5. MSE for outliers.

γ_0	$p=0$	$p=1$	$p=1(N)$	$p=1(S)$	$p=2$
$\phi/2$	1.0021	1.8085	2.8196	0.8188	1.9953
$\phi/4$	1.1022	1.0010	2.8345	0.8883	2.0041
$\phi/2$	1.6233	0.9902	2.8133	0.7085	1.8030
$\phi/4$	1.5033	1.0011	3.0011	0.8355	1.9224

The corresponding execution times are shown in table 6.

Table 6. Time for outliers (sec).

γ_0	$p=0$	$p=1$	$p=1(N)$	$p=1(S)$	$p=2$
$\phi/2$	1.5091	1.5749	7.7632	0.4000	1.5100
$\phi/4$	1.5885	1.5345	8.0123	0.4123	1.5108
$\phi/2$	1.5990	1.5666	8.0990	0.4022	1.5031
$\phi/4$	1.5462	1.5778	8.1435	0.4011	1.5022

Finally, a linear chirp has been added to the original time series with amplitude half of the ECG mean, and start and stop frequency equal to $\pi/8$ and $\pi/2$. The MSE between the regularized and the original waveforms is computed.

Table 7. MSE for linear chirp.

γ_0	$p=0$	$p=1$	$p=1(N)$	$p=1(S)$	$p=2$
$\phi/2$	2.0027	1.9937	4.0002	1.8666	1.8002
$\phi/4$	1.1103	2.0039	6.5835	1.9014	1.8999
$\phi/2$	2.9923	1.9499	8.2400	1.8535	1.4436
$\phi/4$	2.5608	2.0002	8.4562	1.8034	2.0003

The corresponding execution times are shown in table 8.

As a genral remark, the reason regularization performs well is the almost periodic nature of heartbeat rate. By placing constraints on the second derivative of $\mathbf{s}_p^*$, the non-periodic components, including the chirp, are removed or at least smoothed. The latter is espcially true for outliers, since they contribute more to total signal energy. It should be noted that regularization is sensitive to both p and γ_0, with the latter being more important for the specific dataset. Moreover, increasing p does not necessarily reduce MSE, implying it is dataset dependent. Thus a global or local minimizing p must be sought. Total execution time suggests the proposed regularization scheme is lightweight enough to be applied to large data vectors.

Table 8. Time for linear chirp (sec).

γ_0	$p = 0$	$p = 1$	$p = 1(N)$	$p = 1(S)$	$p = 2$
$\phi/2$	3.1146	3.2463	6.8338	0.4100	4.2371
$\phi/4$	3.7734	4.0354	8.0023	0.4053	4.0123
$\phi/2$	4.0453	3.0456	7.8123	0.4618	4.7734
$\phi/4$	4.0228	4.7745	8.9932	0.4995	4.4527

7 Conclusions and Future Work

Finite difference regularization is an intuitive way of impose smoothness constraints on a data vector with a large number of erroneous, missing, or outlying values when a data cleansing model cannot be built. It relies on a tradeoff between the similarity of the regularized data to the original ones and the smoothness of the regularized data. Also, the relationship to the Laplace operator is explored for a special case of difference order. The applicability of the above is demonstrated using ECG biosignals from the MIT-BIH benchmark dataset.

Future research directions are the extension of the cost function to other, possibly non-differentiable, norms as, another norm may offer a higher level of data insight. Finally, distributed methods over the Hadoop ecosystem for machine learning platforms such as Spark can be developed.

Acknowledgements

This work was supported by the European Unions Horizon 2020 research and innovation programme under grant agreement No 690140.

References

1. Acharya, R., Krishnan, S.M., Spaan, J.A., Suri, J.S.: Advances in cardiac signal processing. Springer (2007)
2. Barron, A.R.: Complexity regularization with application to artificial neural networks. In: Nonparametric functional estimation and related topics, pp. 561–576. Springer (1991)
3. Baselli, G., Cerutti, S., Civardi, S., Lombardi, F., Malliani, A., Merri, M., Pagani, M., Rizzo, G.: Heart rate variability signal processing: A quantitative approach as an aid to diagnosis in cardiovascular pathologies. International journal of biomedical computing 20(1), 51–70 (1987)
4. Begum, N., Hu, B., Rakthanmanon, T., Keogh, E.: Towards a minimum description length based stopping criterion for semi-supervised time series classification. In: 14th IEEE International Conference on Information Reuse and Integration (IRI). pp. 333–340. IRI 2013, IEEE (2013)
5. Bishop, C.M.: Pattern recognition. Machine Learning (2006)
6. Calvetti, D., Morigi, S., Reichel, L., Sgallari, F.: Tikhonov regularization and the L-curve for large discrete ill-posed problems. Journal of computational and applied mathematics 123(1), 423–446 (2000)

7. Demmel, J.W.: Applied numerical linear algebra. SIAM (1997)
8. Drakopoulos, G., Baroutiadi, A., Megalooikonomou, V.: Higher order graph centrality measures for Neo4j. In: Proceedings of the 6th International Conference of Information, Intelligence, Systems, and Applications. IISA 2015, IEEE (July 2015)
9. Drakopoulos, G., Megalooikonomou, V.: On the weight sparsity of multilayer perceptrons. In: Proceedings of the 6th International Conference of Information, Intelligence, Systems, and Applications. IISA 2015, IEEE (July 2015)
10. Edmunds, D., Rákosník, J.: Sobolev embeddings with variable exponent. Studia Mathematica 143(3), 267–293 (2000)
11. Fuhry, M., Reichel, L.: A new Tikhonov regularization method. Numerical Algorithms 59(3), 433–445 (2012)
12. Fukumizu, K., Bach, F.R., Jordan, M.I.: Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. The Journal of Machine Learning Research 5, 73–99 (2004)
13. Girosi, F., Jones, M.B., Poggio, T.: Regularization theory and neural networks architectures. Neural computation 7(2), 219–269 (1995)
14. Goldberger, A., Rigney, D.: Nonlinear dynamics at the bedside. In: Glass, L., Hunter, P., McCulloch, A. (eds.) Theory of Heart: Biomechanics, Biophysics, and Nonlinear Dynamics of Cardiac Function, pp. 583–605. Springer-Verlag (1991)
15. Golub, G.H., Hansen, P.C., O'Leary, D.P.: Tikhonov regularization and total least squares. SIAM Journal on Matrix Analysis and Applications 21(1), 185–194 (1999)
16. Guvenir, H.A., Acar, B., Demiroz, G., Cekin, A.: A supervised machine learning algorithm for arrhythmia analysis. In: Proceedings of the Computers in Cardiology conference (1997)
17. Haykin, S.S.: Neural networks and learning machines. Pearson Education (2009)
18. Lee, D.D., Seueng, S.H.: Learning the parts of objects by non-negative matrix factorization. Letters to nature 1(1) (December 1994)
19. Malik, M., Farrell, T., Cripps, T., Camm, A.: Heart rate variability in relation to prognosis after myocardial infarction: selection of optimal processing techniques. European heart journal 10(12), 1060–1074 (1989)
20. Natterer, F.: Error bounds for Tikhonov regularization in Hilbert scales. Applicable Analysis 18(1-2), 29–37 (1984)
21. Ng, A.Y.: Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In: Proceedings of the twenty-first International Cconference on Machine Learning. pp. 78–65. ICML 2004, ACM (2004)
22. O'Leary, D.P.: Near-optimal parameters for Tikhonov and other regularization methods. SIAM Journal on scientific computing 23(4), 1161–1171 (2001)
23. Shewchuk, J.: Conjugate gradient without the agonizing pain. Tech. rep., Carnegie Mellon University (2007)
24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014)
25. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
26. Tikhonov, A., Goncharsky, A., Stepanov, V., Yagola, A.: Numerical methods for the solution of ill-posed problems. Kluwer Academic Publishers (1995)
27. Ye, C., Pallauf, J., Kumar, B., Coimbra, M.T.: Customizing the training dataset to an individual for improved heartbeat recognition performance in long-term ECG signals. In: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. pp. 3322–3325. IEEE (2011)