Data-driven Thresholding in Denoising with Spectral Graph Wavelet Transform

Basile de Loynes, Fabien Navarro, Baptiste Olivier

Data-driven Thresholding in Denoising with Spectral Graph Wavelet Transform

Basile de Loynes*, Fabien Navarro†, Baptiste Oliver‡

July 27, 2020

Abstract

This paper is devoted to adaptive signal denoising in the context of Graph Signal Processing (GSP) using Spectral Graph Wavelet Transform (SGWT). This issue is addressed *via* a data-driven thresholding process in the transformed domain by optimizing the parameters in the sense of the Mean Square Error (MSE) using the Stein's Unbiased Risk Estimator (SURE). The SGWT considered is built upon a partition of unity making the transform semi-orthogonal so that the optimization can be performed in the transformed domain. However, since the SGWT is over-complete, the divergence term in the SURE needs to be computed in the context of correlated noise. Two thresholding strategies called coordinatewise and block thresholding process are investigated. For each of them, the SURE is derived for a whole family of elementary thresholding functions among which the soft threshold and the James-Stein threshold. This multi-scales analysis shows better performance than the most recent methods from the literature. That is illustrated numerically for a series of signals on different graphs.

1 Introduction

The emerging field of Graph Signal Processing (GSP) aims to bridge the gap between signal processing and spectral graph theory (see for instance Chung (1997); Belkin and Niyogi (2008) and references therein). One objective is to generalize fundamental analysis operations from regular grid signals to irregular structures as graphs. There is an extensive literature on GSP, in particular we refer the reader to Shuman et al. (2013) for an introduction to this field and Ortega et al. (2018) for an overview of recent developments, challenges and applications. As a matter of fact, GSP have already been applied in machine/deep learning: convolutional neural networks (CNN) on graphs Bruna et al. (2014); Henaff et al. (2015); Defferrard et al. (2016), semi-supervised classification with graph CNN Kipf and Welling (2017); Hamilton et al. (2017), community detection Tremblay and Borgnat (2014), to name just a few. In the context of GSP, the authors of Coifman and Lafon (2006); Gavish et al. (2010); Hammond et al. (2011) have developed wavelet transforms on graphs. More specifically, in Hammond et al. (2011) a fairly general construction of a frame enjoying the usual properties of standard wavelets is developed: each vector of the frame is localized both in the graph domain and the spectral domain. The transform associated with this frame is named Spectral Graph Wavelet Transform (SGWT). Many studies based on SGWT (or some variants) explore the denoising performance of this

*Basile de Loynes
ENSAI, France, E-mail: basile.deloynes@ensai.fr
†Fabien Navarro
CREST, ENSAI, France, E-mail: fabien.navarro@ensai.fr
‡Baptiste Oliver
Orange Labs, France. E-mail: baptiste.olivier@orange.fr

approach using different strategies Leonardi and Van De Ville (2013); Onuki et al. (2016); Wang et al. (2016); Deutsch et al. (2016); Irion and Saito (2017); Dong et al. (2016); Göbel et al. (2018) from signal adapted tight frames to regularization method.

The denoising approach chosen in this paper involves several thresholding processes in the transformed domain of the wavelet coefficients. Actually, this can be seen as an extension to SGWT of the methodology of Donoho and Johnstone (1995); Cai (1999). With this approach, the main challenge is the efficient calibration of the parameters minimizing the MSE risk in a complete data-driven way. Recently, in the setting of discrete wavelets transform on a regular grid—the so-called regular case—the Stein's unbiased risk estimate (SURE) has proven to be a powerful tool for signal/image restoration Luisier et al. (2007); Pesquet et al. (2009); Vaiter et al. (2013). Based on the Stein's lemma, this estimator acts as a proxy for the MSE which cannot be computed in practice since the original signal is unknown. In this paper, the SURE is explicitly computed for an arbitrary thresholding process in Theorem 1 and for correlated noise in the graph domain in Corollary 1. Also, let us point out that contrary to the regular wavelet transform, the SGWT is no longer orthogonal so that a white Gaussian noise in the graph domain is transformed in a correlated noise. Consequently, the divergence term of the resulting SURE involves the covariance of the transformed noise making the numerical evaluation less simple than in the regular case. Afterward, the SURE is specified to the case of coordinatewise and block thresholding. The latter is inspired by image denoising problems for which a Stein risk estimator has been proposed in Peyré et al. (2011) to tune both the block-sparsity structure and the threshold. A similar selection strategy has been developed by Navarro et al. (2013) in the context of deconvolution. The R package **gasper** which implements the method introduced in this paper is available on github[1] (de Loynes et al., 2020) as well as the scripts to reproduce the results presented[2].

Regarding the regularization method implemented in Onuki et al. (2016), the regularization parameter is also selected optimizing an MSE proxy based on a similar argument. Nonetheless, beyond the fact that the philosophy is different (regularization *versus* thresholding), one stress that the empirical risk bias is explicitly determined while the MSE estimation in Onuki et al. (2016) is only validated numerically. Another penalization method is given in Wang et al. (2016) that extends the approach from Tibshirani and Taylor (2011) within the framework of graphs. For this method, the divergence term is computed explicitly; this gives rise to a data-driven parameter selection method so that this approach is an interesting concurrent to our methodology.

The paper is organized as follows. Section 2 introduces the notation and briefly reviews the notions of tight frame and SGWT of Hammond et al. (2011). Section 3 is devoted to denoising and the SURE estimator for generic thresholding process in the transformed domain. Then, the SURE is specified in the cases of coordinatewise, block thresholding processes and for correlated noise in the graph domain. In Section 4 numerical comparisons with the classical Wiener filter (oracle version) and the trend filtering introduced in Wang et al. (2016) for denoising are discussed. Several signals and graphs, including examples from real datasets, are considered. For these experimental results, the construction of the frame follows Göbel et al. (2018). In terms of denoising performance, other tight frames such as spectrum adapted and/or signal adapted tight frames from Shuman et al. (2015) and Behjat et al. (2016) might give better results. Still, Theorem 1 actually applies to any tight frame and the question of exhibiting the most efficient one is beyond the scope of the paper.

[1] https://github.com/fabnavarro/gasper
[2] https://github.com/fabnavarro/SGWT-SURE

2 Spectral Graph Wavelet Transform

2.1 Graphs, Frames and Tight Frames

Let G be an undirected weighted graph, with set of vertices V, and weights $(w_{ij})_{i,j\in V}$ satisfying $w_{ij} = w_{ji}$ for $i, j \in V$. The size of the graph is the number of nodes $n = |V|$. The (unnormalized) graph Laplacian matrix $\mathcal{L} \in \mathbb{R}^{V\times V}$ associated with G is the symmetric matrix defined as $\mathcal{L} = D - W$, where W is the matrix of weights with coefficients $(w_{ij})_{i,j\in V}$, and D the diagonal matrix with diagonal coefficients $D_{ii} = \sum_{j\in V} w_{ij}$. A signal f on the graph G is a function $f : V \to \mathbb{R}$.

Let $\mathfrak{F} = \{r_i\}_{i\in I}$ be a frame of vectors of $\mathbb{R}^V$, that is a family of vectors in $\mathbb{R}^V$ such that there exist $A, B > 0$ satisfying for all $f \in \mathbb{R}^V$

$$A\|f\|_2^2 \leq \sum_{i\in I} |\langle f, r_i\rangle|^2 \leq B\|f\|_2^2. \tag{1}$$

The linear map $T_{\mathfrak{F}} : \mathbb{R}^V \to \mathbb{R}^I$ defined for $f \in \mathbb{R}^V$ by $T_{\mathfrak{F}} f = (\langle f, r_i\rangle)_{i\in I}$ is called the *analysis* operator. The *synthesis* operator is the adjoint of $T_{\mathfrak{F}}$: namely, it is the linear map $T_{\mathfrak{F}}^* : \mathbb{R}^I \to \mathbb{R}^V$ defined for a vector of coefficients $(c_i)_{i\in I}$ by $T_{\mathfrak{F}}^*(c_i)_{i\in I} = \sum_{i\in I} c_i r_i$. As a frame is in particular a generating family of $\mathbb{R}^V$, a signal $f \in \mathbb{R}^V$ can be recovered from its coefficients $T_{\mathfrak{F}} f$ with the help of the synthesis operator.

2.2 Construction of Tight Frames

A frame $\mathfrak{F}$ is said to be tight if $A = B = 1$ in Equation (1)—the latter is then termed the Parseval identity. From now on, the frames considered are supposed to be tight. Let us recall the generic construction of such a frame (*c.f.* Kereta et al. (2019) for instance).

Since $\mathcal{L}$ is self-adjoint, it admits the spectral decomposition $\mathcal{L} = \sum_\ell \lambda_\ell \langle \chi_\ell, \cdot\rangle \chi_\ell$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = 0$ denote the (ordered) eigenvalues of the matrix $\mathcal{L}$, and $(\chi_\ell)_{1\leq \ell\leq n}$ are the associated normalized and pairwise orthogonal eigenvectors. Then, for any function $\rho : \mathrm{sp}(\mathcal{L}) \to \mathbb{R}$ defined on the spectrum $\mathrm{sp}(\mathcal{L})$ of matrix $\mathcal{L}$, the functional calculus formula reads $\rho(\mathcal{L}) = \sum_\ell \rho(\lambda_\ell)\langle \chi_\ell, \cdot\rangle \chi_\ell$. A finite collection $(\psi_j)_{j=0,\dots,J}$ is a finite partition of unity on the compact $[0, \lambda_1]$ if $\psi_j : [0, \lambda_1] \to [0, 1]$ for all $j \in \mathcal{J}$ and $\forall \lambda \in [0, \lambda_1]$, $\sum_{j=0}^J \psi_j(\lambda) = 1$. Given a finite partition of unity $(\psi_j)_{j=0,\dots,J}$, the Parseval identity implies that the following set of vectors is a tight frame:

$$\mathfrak{F} = \left\{ \sqrt{\psi_j}(\mathcal{L})\delta_i, j = 0, \dots, J, i \in V \right\}.$$

Also, following Leonardi and Van De Ville (2013); Göbel et al. (2018), a partition of unity can be easily defined as follows: let $\omega : \mathbb{R}^+ \to [0, 1]$ be some function with support in $[0, 1]$, satisfying $\omega \equiv 1$ on $[0, b^{-1}]$, for some $b > 1$, and set

$$\psi_0(x) = \omega(x) \quad \text{and} \quad \psi_j(x) = \omega(b^{-j}x) - \omega(b^{-j+1}x) \quad \text{for} \quad j = 1, \dots, J, \quad \text{where} \quad J = \left\lfloor \frac{\log \lambda_1}{\log b} \right\rfloor + 2.$$

2.3 Discrete SGWT Associated with a Partition of Unity

Let $(\psi_j)_{j=0,\dots,J}$ be a partition of unity of $[0, \lambda_1]$. The SGWT of a signal $f \in \mathbb{R}^V$ is given by

$$\mathcal{W} f = \left(\sqrt{\psi_0}(\mathcal{L}) f^T, \dots, \sqrt{\psi_J}(\mathcal{L}) f^T \right)^T \in \mathbb{R}^{n(J+1)}.$$

The adjoint linear transformation $\mathcal{W}^*$ of $\mathcal{W}$ is:

$$\mathcal{W}^* \left(\eta_0^T, \eta_1^T, \dots, \eta_J^T \right)^T = \sum_{j\geq 0} \sqrt{\psi_j}(\mathcal{L})\eta_j.$$

The tightness of the underlying frame implies that $\mathcal{W}^*\mathcal{W} = \mathrm{Id}_{\mathbb{R}^V}$ so that a signal $f \in \mathbb{R}^V$ can be recovered by applying $\mathcal{W}^*$ to its wavelet coefficients $((\mathcal{W} f)_i)_{i=1,\dots,n(J+1)} \in \mathbb{R}^{n(J+1)}$ (see Hammond et al. (2011)).

3 Adaptive Denoising with SGWT

Let $f \in \mathbb{R}^V$ be some signal on a graph G and ξ be an n-dimensional Gaussian vector distributed as $\mathcal{N}(0, \sigma^2 \mathrm{Id})$. The aim of denoising is to recover the unknown signal f from the observed noisy version $\tilde{f} = f + \xi$. Basically, our denoising procedure will consist of three steps: (1) compute the SGWT transform $\mathcal{W}\tilde{f} \in \mathbb{R}^{n(J+1)}$; (2) apply a given thresholding operator $h(\cdot)$ to the coefficients $\mathcal{W}\tilde{f}$; (3) apply the inverse SGWT transform to obtain an estimation $\hat{f}$ of the original signal. Here, the main challenge in denoising consists in choosing a suitable thresholding operator with respect to the noisy signal $\tilde{f}$ and the underlying graph. The performance measure in the sequel will be the MSE between the original signal f and the denoised signal $\hat{f}$: $\|f - \hat{f}\|_2^2$. First, it is worth noting that the Parseval identity allows direct optimization in the transformed domain of wavelet coefficients. Secondly, in practice, obviously the original signal remains unknown. To overcome this difficulty, the MSE is generally substituted with the Stein's Unbiased Risk Estimator which no longer depends on the original signals (see Donoho and Johnstone (1995) for instance). Nonetheless, contrary to the usual wavelet transform, the white noise ξ is mapped onto a correlated Gaussian noise. In the next section, the SURE is derived taking into account these correlations.

3.1 The SURE Estimator in the Transformed Domain

By linearity, the denoising problem $\widetilde{f} = f + \xi$ is transferred to the denoising problem $\widetilde{F} = F + \Xi$ with $\Xi \sim \mathcal{N}(0, \sigma^2 \mathcal{W}\mathcal{W}^*)$, $\widetilde{F} = \mathcal{W}\widetilde{f}$ and $F = \mathcal{W}f$. The spectral decomposition of $\mathcal{W}\mathcal{W}^*$ reads $\mathcal{W}\mathcal{W}^* = U\Sigma U^*$ with U a unitary matrix of $\mathbb{R}^{n(J+1)}$ and $\Sigma = \left(\begin{smallmatrix} \mathrm{Id}_{\mathbb{R}^n} & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

A thresholding process is a map $h : \mathbb{R}^{n(J+1)} \to \mathbb{R}^{n(J+1)}$. Typically, the map h is a coordinate-wise or a block shrinkage in applications. The following result extending the SURE's expression to correlated noise is based on the Stein's lemma in Stein (1981) in which h is assumed to be weakly differentiable. One refers the reader to Stein (1981) for the precise definition.

Theorem 1 (h-SURE)**.** *Let h be a weakly differentiable thresholding process for the denoising problem $\widetilde{F} = F + \Xi$. Then the theoretical MSE is given by*

$$\mathbf{E}\|h(\widetilde{F}) - F\|^2 = \mathbf{E}\left[-n\sigma^2 + \|h(\widetilde{F}) - \widetilde{F}\|^2 + 2\sum_{i,j=1}^{n(J+1)} \mathbf{Cov}(\Xi_i, \Xi_j)\partial_j h_i(\widetilde{F})\right],$$

where h_i is the i-th component of h.

It is worth noting that $\mathbf{Cov}(\Xi_i, \Xi_j) = \sigma^2(\mathcal{W}\mathcal{W}^*)_{i,j}$ so that, as soon as the thresholding process h is specified and the noise variance σ^2 estimated, the SURE of h defined below can be completely computed from the noisy observations as in the regular case:

$$\mathbf{SURE}(h) = -n\sigma^2 + \|h(\widetilde{F}) - \widetilde{F}\|^2 + 2\sum_{i,j=1}^{n(J+1)} \mathbf{Cov}(\Xi_i, \Xi_j)\partial_j h_i(\widetilde{F}).$$

Proof. The theoretical MSE can be rewritten as follows

$$\mathbf{E}\|h(\widetilde{F}) - F\|^2 = \mathbf{E}\|h(\widetilde{F}) - \widetilde{F}\|^2 + \mathbf{E}\|\Xi\|^2 + 2\mathbf{E}\langle h(\widetilde{F}) - \widetilde{F}, \Xi\rangle.$$

The second term is equal to $n\sigma^2$ since almost surely $\|\Xi\|^2 = \|U^*\Xi\|^2 = \|P_K U^*\Xi\|^2$ where $K = \ker(\mathcal{W}\mathcal{W}^*)^\perp$ and P_K the orthogonal projection onto K. Finally, setting $g(x) = h(x) - x$, $x \in \mathbb{R}^{n(J+1)}$, it remains to compute the last term $\mathbf{E}\langle g(\widetilde{F}), \Xi\rangle = \mathbf{E}\langle g(F + \Xi), \Xi\rangle$ where F is deterministic and $\Xi \sim \mathcal{N}(0, \sigma^2 \mathcal{W}\mathcal{W}^*)$. A simple computation gives

$$\mathbf{E}\langle g(F + \Xi), \Xi\rangle = \sum_{i=1}^{n(J+1)} \mathbf{E}[g_i(F + \Xi)\Xi_i] = \sum_{i=1}^{n(J+1)} \mathbf{Cov}(g_i(F + \Xi), \Xi_i).$$

Then, following Liu (1994), each term in the sum above is given by

$$\mathbf{Cov}(g_i(F+\Xi),\Xi_i) = -n\sigma^2 + \sum_{j=1}^{n(J+1)} \mathbf{Cov}(\Xi_i,\Xi_j)\mathbf{E}[(\partial_j h_i)(F+\Xi)],$$

since $\mathrm{Tr}(\sigma^2 \mathcal{W}\mathcal{W}^*) = n\sigma^2$. This ends the proof of Theorem 1. □

3.2 Coordinatewise Thresholding Process

For a coordinatewise thresholding process, the map h is of the form $h(x) = (\tau(x_i,t_i))_{i=1,\dots,n(J+1)}$ where $(t_i)_{i=1,\dots,n(J+1)}$ are the thresholds. In practice, we may choose $\tau(x,t) = x\max\{1 - t^\beta |x|^{-\beta}, 0\}$ with $\beta \geq 1$. The most popular choices are the soft thresholding ($\beta = 1$), the James-Stein thresholding ($\beta = 2$) and the hard thresholding ($\beta = \infty$). The latter will not be considered here since it does not lead to a sufficiently regular thresholding process for Theorem 1 to be applied.

For any $\beta \in [1,\infty)$, the derivative $\partial_j h_i$ vanishes whereas

$$\partial_i h_i(F+\Xi) = \mathbf{1}_{[t_i,\infty)}(|\widetilde{F}_i|)\left[1 + (\beta-1)\frac{t^\beta}{|\widetilde{F}_i|^\beta}\right].$$

Consequently, the SURE associated with h is given by

$$\mathbf{SURE}(h) = -n\sigma^2 + \sum_{i=1}^{n(J+1)} \widetilde{F}_i^2 \left(1 \wedge \frac{t_i^\beta}{|\widetilde{F}_i|^\beta}\right)^2 + 2\sum_{i=1}^{n(J+1)} \mathbf{V}(\Xi_i)\mathbf{1}_{[t_i,\infty)}(|\widetilde{F}_i|)\left[1 + \frac{(\beta-1)t_i^\beta}{|\widetilde{F}_i|^\beta}\right]. \quad (2)$$

The usual expression of the SURE is recovered from the identity above remarking that $\mathbf{V}[\Xi_i]$ are identically equal to σ^2 when the transformed noise is uncorrelated.

Let us notice that the coordinatewise soft thresholding ($\beta = 1$) satisfies an oracle inequality as shown in Göbel et al. (2018). Similarly to the regular case, it states that up to a log factor, the soft thresholding estimator can mimic an oracle projection.

3.3 Optimization: Donoho and Johnstone's Trick

The SURE can be optimized in the same way as in the standard case using the Donoho and Johnstone's trick of Donoho and Johnstone (1995) whose the justification is recalled below.

For the sake of simplicity, we first consider the case of the coordinatewise thresholding process with a uniform threshold: $t_i = t$ for all $i = 1,\dots,n(J+1)$. Denote by $a_1,\dots,a_{n(J+1)}$ the absolute values of the noisy wavelet coefficients $|\widetilde{F}_i|$ in the increasing order. The trick comes from the observation that, on each interval (a_k, a_{k+1}), the last term of Equation (2) is non-decreasing whereas the second term is an increasing function of t. Consequently, the SURE hits its minimum at some value a_{k^*}, $k^* = 1,\dots,n(J+1)$.

If the thresholds t_i are no longer uniform but merely tied inside blocks with values $t_1,\dots,t_L$, the same trick is still valid: group the terms in the sums along the different parameters $t_1,\dots,t_L$ and optimize each partial sum with respect to t_k, $k = 1,\dots,L$.

3.4 Block Thresholding Process

In order to take advantage of the localization properties of SGWT and the regularity of the original signal, we may introduce block thesholding processes similar to Cai (1999).

Consider a partition $(B_\ell)_{\ell\in L}$ of $\{1,\dots,n(J+1)\}$ and set $\|x\|_{B_\ell}^2 = \sum_{i\in B_\ell}(x_i)^2$. In this case, the thesholding process $h = (h_i)_{i=1,\dots,n(J+1)}$ reads

$$h_i(x) = x_i \max\left\{1 - \frac{t_\ell^\beta}{\|x\|_{B_\ell}^\beta}, 0\right\}, \quad x \in \mathbb{R}^{n(J+1)}, \quad \text{and} \quad \ell \in L : i \in B_\ell.$$

If i, j are in different blocks, then $\partial_j h_i$ vanishes. Additionally, if i, j are in B_ℓ but $i \neq j$ then

$$\partial_j h_i(\widetilde{F}) = \widetilde{F}_i \mathbf{1}_{[t_\ell,\infty)}(\|\widetilde{F}\|_{B_\ell}) \beta t_\ell^\beta \widetilde{F}_j \|\widetilde{F}\|_{B_\ell}^{-\beta-2},$$

whereas

$$\partial_i h_i(\widetilde{F}) = \mathbf{1}_{[t_\ell,\infty)}(\|\widetilde{F}\|_{B_\ell}) \left(1 - t_\ell^\beta \|\widetilde{F}\|_{B_\ell}^{-\beta} + \beta t_\ell^\beta \widetilde{F}_i^2 \|\widetilde{F}\|_{B_\ell}^{-\beta-2}\right).$$

Consequently, a straightforward computation leads to

$$\begin{aligned}\mathbf{SURE}(h) = -n\sigma^2 &+ \sum_{\ell\in L} \left(1 \wedge \frac{t_\ell^\beta}{\|\widetilde{F}\|_{B_\ell}^\beta}\right)^2 \|\widetilde{F}\|_{B_\ell}^2 \\ &+ 2\sum_{\ell\in L} \mathbf{1}_{[t_\ell,\infty)}(\|\widetilde{F}\|_{B_\ell}) \left[\left(1 - \frac{t_\ell^\beta}{\|\widetilde{F}\|_{B_\ell}^\beta}\right) \sum_{i\in B_\ell} \mathbf{V}(\Xi_i) + \frac{\beta t_\ell^\beta}{\|\widetilde{F}\|_{B_\ell}^{\beta+2}} \sum_{i,j\in B_\ell} \mathbf{Cov}(\Xi_i, \Xi_j)\widetilde{F}_i\widetilde{F}_j\right]\end{aligned}$$

Once again, for uncorrelated transformed noise, the usual expression easily follows from the identity above. Note also that the optimization of the SURE in this case requires more sophisticated techniques as the divergence term is no longer monotone.

3.5 Correlated Noise in the Graph Domain

The SURE can also be stated in the context of correlated noise at the cost of some prior information on the covariance structure. More precisely, in the denoising problem $\widetilde{f} = f + \xi$ with correlated noise, it is supposed that $\xi \sim \mathcal{N}(0, \Gamma)$ for some covariance matrix Γ. The denoising problem reads in the transformed problem as $\widetilde{F} = F + \Xi$ with $\Xi \sim \mathcal{N}(0, \mathcal{W}\Gamma\mathcal{W}^*)$.

Corollary 1. *Under the assumption of Theorem 1, the theoretical* MSE *is given by*

$$\mathbf{E}[\|h(\widetilde{F}) - F\|^2] = \mathbf{E}\left[-\operatorname{Tr}(\mathcal{W}\Gamma\mathcal{W}^*) + \|h(\widetilde{F}) - \widetilde{F}\|^2 + 2\sum_{i,j=1}^{n(J+1)} \mathbf{Cov}(\Xi_i, \Xi_j)\partial_j h_i(\widetilde{F})\right],$$

where h_i is the i-th component of h.

Let us point out that the parameters selection can be made without computing explicitly $\operatorname{Tr}(\mathcal{W}\Gamma\mathcal{W}^*)$ since it does not depend on h—even though, the MSE estimate is obviously shifted by this quantity. Besides, the correlation structure Γ is actually hidden in the quantities $\mathbf{Cov}(\Xi_i, \Xi_j)$, namely, for $1 \leq i, j \leq n(J+1)$: $\mathbf{Cov}(\Xi_i, \Xi_j) = \left(\mathcal{W}\Gamma\mathcal{W}^*\right)_{i,j}$. Consequently, computationally speaking, there is no additional burden compared to the white noise case.

Proof. The proof follows the lines of Theorem 1 with $\mathbf{E}[\Xi_i^2] = \left(\mathcal{W}\Gamma\mathcal{W}^*\right)_{i,i}, 1 \leq i \leq n(J+1)$. □

In applications, it is usually reasonable to assume some structure on the covariance matrix Γ reflecting the topology of the underlying graph. Typically, the noise on two given vertices may be correlated if those vertices are close enough in the graph. For example, let $\xi_0 \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$, we set $\xi = \xi_0 + \alpha W\xi_0$ where W is the graph matrix of weights and $\alpha > 0$ some tuning parameter describing the global intensity of the correlation. Then, it follows,

$$\Gamma = \sigma^2(\mathrm{Id} + 2\alpha W + \alpha^2 WW^*). \tag{3}$$

Other choices are obviously possible.

3.6 Complexity

Regarding the space complexity, we need to store the frame and the weights appearing in the SURE for a cost of $O(n^2(J+1)^2)$. With given Laplacian eigendecomposition, the time complexity of the optimization of the SURE is (in average) of order $O(n(J+1)\log(n(J+1)))$ for the coordinate-wise estimator following [8]. For the block estimator, the use of a grid search is a limitation.

4 Numerical Results

This section presents the empirical performance of the proposed automatic threshold selection for different signals defined on different graphs: the Minnesota roads graph (seen as a reference in many recent studies, see Behjat et al. (2016) and references therein) with synthetic signals and the Facebook graph with signals from Wang et al. (2016), the Pittsburgh Census Tract graph, a graph built from a dataset on New York City taxis with a real signal as well as numerical experiments in the correlated and block cases. All the experiments are conducted with the R package **gasper**.

4.1 The Minnesota Roads Graph

The Minnesota roads graph is a planar graph consisting of 2642 vertices and 6606 edges. Each vertex is described by its (x, y)-coordinates. The function ω chosen in the experiments is a piecewise linear function with support in $[0, 1]$ and constant equal to 1 on $[0, b^{-1}]$ with $b = 2$. From $\lambda_1 \approx 6.89$, we deduce that the number of scales is $J + 1 = 5$—see Section 2.

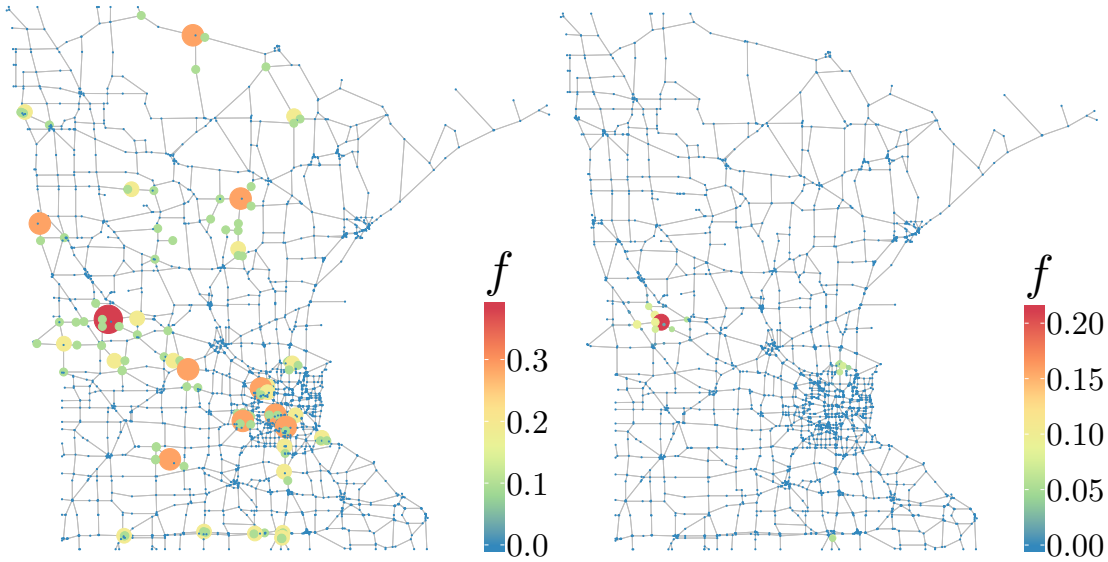


Figure 1: Signals used on the Minnesota graph.

On this graph, two classes of synthetic signals are generated inspired by the methodology introduced in Behjat et al. (2016). Let us briefly recall the construction: let $\eta \in (0, 1)$ and $k \in \mathbb{N}$ be two parameters; a signal $f_{\eta,k}$ is obtained by letting the adjacency matrix W acts on an *i.i.d.* realization x_η of Bernoulli random variables of parameter η, in symbols $f_{\eta,k} = W^k x_\eta / \lambda_1^k$. This method generates signals with different regularities. This has to be understood in the sense of the graph topology and not that one given by the embedding space $\mathbb{R}^2$. In the experiment, two signals are generated with parameters $\eta = 0.01, k = 2$ and $\eta = 0.001, k = 4$ respectively (see Figure 1 left and right respectively).

We compare the performance in terms of SNR (computed on the functions after reconstruction) for different denoising strategies, different noise levels and each synthetic signal $f_{\eta,k}$. For each noise level $\sigma = 0.005, 0.01, 0.02$ and $\sigma = 0.001, 0.002, 0.004$, a sample of $N = 10$ white Gaussian noise is simulated and a global (G) *versus* level-dependent (LD) coordinatewise thresholding are performed with respect to the soft ($\beta = 1$) and James-Stein ($\beta = 2$) thresholding rules. For each strategy, we compare the average behavior of the SNR for parameter selected with the oracle ($\mathrm{MSE}^{\beta=1,2}$ obtained by minimizing the MSE using the original signal f) and the SURE with known σ ($\mathrm{SURE}_{\sigma}^{\beta=1,2}$). Also, the standard deviation on the sample is provided.

These results are first compared to the classical Wiener filter. More precisely, the Wiener filter consists of attenuating the Fourier coefficient $\mathcal{F}(\widetilde{f})$ of $\widetilde{f}$. Below, we only consider the oracle linear attenuation $\mathcal{F}(\widetilde{f})[i]\mathcal{F}(f)[i]/(\mathcal{F}(f)[i]^2 + \sigma^2)$. While this estimator is unrealistic since it depends on f, any Wiener filter has worse performance than this oracle. Table 1 is completed by the performance of the Wiener filter on each signal. Also, the theoretical value r_{inf} of the oracle risk given in Mallat (2009) is recalled for comparison purpose.

Our methodology is also compared to the so-called graph trend filtering (*i.e.* for $k = 0, 1, 2$) introduced in Wang et al. (2016). The graph trend filtering is a regularization method with a penalty term involving the graph difference operator at a given order (see Wang et al. (2016)). In the experiments, we make use of the matlab toolbox **gtf**[3] provided by the authors of Wang et al. (2016).

Table 1: Mean SNR performance over $N = 10$ realizations of the low to high noise levels settings with corresponding empirical standard deviation. Left panel: $f_{0.01,2}$ and right panel: $f_{0.001,4}$.

$\mathrm{SNR}_{\mathrm{in}}$	16.07±0.13	10.05±0.13	4.03±0.13	16.64±0.13	10.62±0.13	4.60±0.13
$\mathrm{MSE}^{\beta=1,\mathrm{G}}$	19.04±0.24	14.22±0.26	9.46±0.26	24.67±0.33	19.77±0.36	14.79±0.45
$\mathrm{MSE}^{\beta=2,\mathrm{G}}$	20.07±0.24	15.60±0.30	10.69±0.30	26.88±0.29	22.18±0.37	17.08±0.67
$\mathrm{SURE}_{\sigma}^{\beta=1,\mathrm{G}}$	18.96±0.27	14.16±0.29	9.46±0.26	24.62±0.37	19.64±0.48	14.70±0.52
$\mathrm{SURE}_{\sigma}^{\beta=2,\mathrm{G}}$	20.04±0.37	15.49±0.43	10.64±0.32	26.73±0.32	21.91±0.52	16.88±0.59
$\mathrm{MSE}^{\beta=1,\mathrm{LD}}$	19.10±0.24	14.28±0.27	9.58±0.27	24.68±0.34	19.79±0.36	14.83±0.46
$\mathrm{MSE}^{\beta=2,\mathrm{LD}}$	20.08±0.24	15.61±0.29	10.72±0.30	26.90±0.26	22.20±0.36	17.13±0.67
$\mathrm{SURE}_{\sigma}^{\beta=1,\mathrm{LD}}$	19.10±0.24	14.26±0.26	9.48±0.24	24.51±0.40	19.59±0.46	14.69±0.49
$\mathrm{SURE}_{\sigma}^{\beta=2,\mathrm{LD}}$	20.01±0.31	15.51±0.36	10.61±0.39	26.52±0.32	21.79±0.34	16.73±0.60
Wiener	17.01±0.13	11.87±0.15	7.43±0.16	17.91±0.12	12.86±0.13	8.40±0.16
r_{inf}	17.05±0.00	11.89±0.00	7.42±0.00	17.96±0.00	12.91±0.00	8.46±0.00
$\mathrm{MSE}^{k=2}$	17.35±0.13	11.43± 0.16	5.68±0.17	19.37±0.14	13.65±0.18	8.01±0.24
$\mathrm{MSE}^{k=1}$	18.05±0.14	11.98±0.18	6.24±0.18	20.43±0.15	14.78±0.18	9.30±0.27
$\mathrm{MSE}^{k=0}$	19.57±0.17	13.43±0.23	7.62±0.22	23.38±0.25	17.88±0.27	12.80±0.40

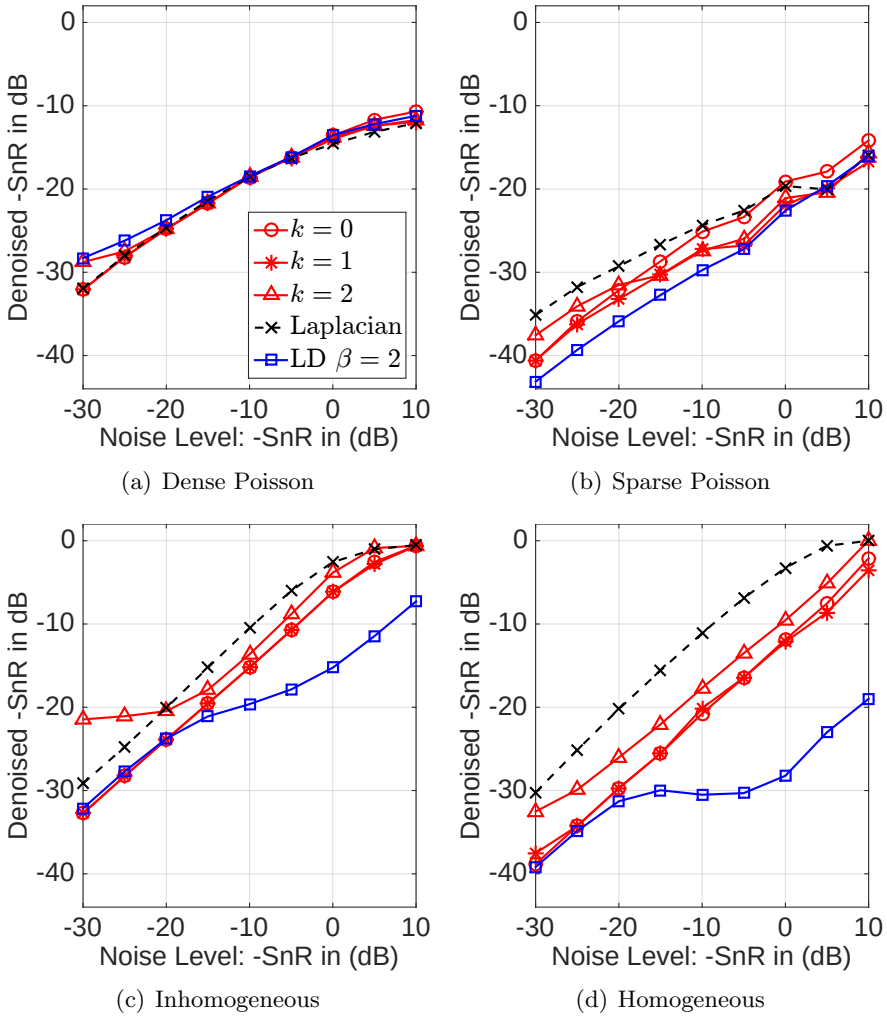
Generally speaking, we observe from Table 1 that our method performs better than the trend filtering motivating the use of multiscale analysis. This idea is confirmed by the comparison with the Wiener filter, in particular in the lower SNR regime that is for higher noise levels. Also, similarly to the regular case, numerical experiments shows that the James-Stein threshold ($\beta = 2$) is slightly more efficient than the soft threshold in particular for the global thresholding process.

Let us point out that there is no fundamental difference in terms of performance between the global and level dependent thresholding in this experiment. In fact, the level dependent thresholding always performs at least as good as the global one. Since the additional computational cost is acceptable, the level dependent thresholding appears to be a good choice without any further *a priori* knowledge.

[3]Available here: https://sites.cs.ucsb.edu/ yuxiangw/resources.html

4.2 The Facebook Graph

Here we examined and compare the denoising performance of level dependent SGWT thresholding (LD) against the trend filtering and Laplacian smoothing Smola and Kondor (2003) on a nonplanar graph considered in Wang et al. (2016): the Facebook graph from the Stanford Network Analysis Project[4]. This undirected graph, collected from survey participants using this Facebook app, is composed of 4039 nodes representing Facebook users, and 88,234 edges representing friendships (see Leskovec and Mcauley (2012) for more details). For signal f, we consider the different regularities used in Wang et al. (2016) as well as the same noise levels, for 5 realizations (see (Wang et al., 2016, Section 5.1) for more details). More precisely, we simply run the scripts provided in the matlab toolbox provided by the authors of Wang et al. (2016). The results are shown in the Figure 2 (to be compared with (Wang et al., 2016, Figure 5 p.14 and Figure 9 p.27)). With the exception of the dense Poisson case, where all methods provide



(a) Dense Poisson (b) Sparse Poisson

(c) Inhomogeneous (d) Homogeneous

Figure 2: Mean SNR performance over $N = 5$ realizations on the Facebook graph.

comparable results, LD globally provides better performance than trend filtering (whatever the value of k), especially at high noise levels where the maximal gain in terms of SNR is greater than 15dB and exceeds 5dB and 10dB respectively for the 5 highest noise levels for inhomogeneous and homogeneous random walk cases. For this graph, the CPU times associated with trend filtering (for $k = 1$ and $k = 2$) for one type of signal and 5 realizations (and a 51-point

[4] http://snap.stanford.edu/data/ego-Facebook.html

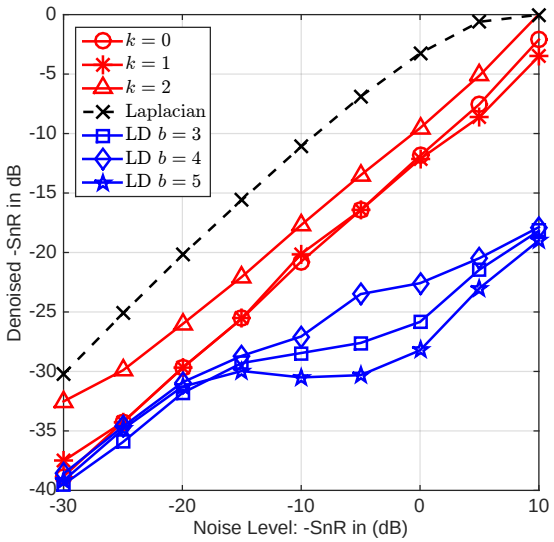


Figure 3: Influence of the number of scales on the denoising performance.

grid search) are of the order of 3 to 4 days (depending on the case), for LD around 25 minutes (including diagonalization and frame calculation which only need to be calculated once).

The calibration of certain parameters has not been studied. However, the latter can considerably influence the performance of the SGWT. For the homogeneous random walk case, we examine here the influence of the number of scales retained for the construction of the frame and controller by b. For $b = 5, 4, 3$, the frame contains respectively $7, 8, 9$ scales. The results are shown in Figure 2, where it can be seen that the frame composed of 7 scales produces the best results. Note that these performances might be improved, for example by making the SURE depend on the β parameter characterizing the threshold rule. The frame considered here has only one parameter, other more flexible constructions, based on a partition of the unit or other types of tight frames such as spectrum adapted and/or signal adapted tight frames from Shuman et al. (2015) and Behjat et al. (2016) could also lead to an improvement.

4.3 Pittsburgh Census Tract Graph

For the sake of completeness, our methodology is also compared to the trend filtering on the Pittsburgh Census Tract graph considered in Wang et al. (2016) which consists of 402 vertices and 2382 edges. The very same piecewise linear function ω with $b = 2$ is still used for this graph. The number of scales is then $J + 1 = 7$. For this experiment, only the level dependent threshold procedure is considered.

We consider the signal and 10 realizations of the noisy signal generated in Wang et al. (2016) (corresponding to an average noise level of 4.84 ± 0.37dB). The resulting SNR for the oracles of SGWT and fused lasso (*i.e.* $k = 0$) are respectively 11.51 ± 0.52dB and 9.85 ± 0.54dB.

Additionally, we run a comparison with the trend filtering (*i.e.* for $k = 0, 1, 2$) for the signal $f_{\eta,k}$ with $\eta = 0.01$ and $k = 5$ with the different noise levels $\sigma = 0.004$, $\sigma = 0.005$ and $\sigma = 0.01$. A comparison with another wavelet estimator proposed in Sharpnack et al. (2013) is also provided, considering two thresholding rules (*i.e.* "soft" and "hard"). For these 5 competitors we only report the oracles results.

Even though the SURE no longer depends on the original signal, it does depend on σ^2 in both methodologies. Since in real applications, the noise level remains unknown in general, we introduce two naive estimators of σ. In fact, a straightforward computation shows that for any

function $g : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\mathbf{E}[\widetilde{f}^T g(\mathcal{L})\widetilde{f}] = f^T g(\mathcal{L}) f + \mathbf{E}[\xi^T g(\mathcal{L})\xi] = f^T g(\mathcal{L}) f + \sigma^2 \mathrm{Tr}\ g(\mathcal{L}),$$

so that a biased estimator of σ^2 is given by

$$\hat{\sigma}_1^2 = \frac{\widetilde{f}^T g(\mathcal{L})\widetilde{f}}{\mathrm{Tr}\ g(\mathcal{L})}.$$

As soon as the original signal is reasonably smooth so that $f^T g(\mathcal{L}) f$ is negligible compared to Tr $g(\mathcal{L})$, then $\hat{\sigma}^2$ is an accurate enough estimation of σ^2. As a first choice, we choose $g(x) = x$. Thanks to Dirichlet's formula, it follows:

$$\hat{\sigma}_1^2 = \frac{\widetilde{f}^T \mathcal{L} \widetilde{f}}{\mathrm{Tr}\ \mathcal{L}} = \frac{\sum_{i,j \in V} w_{ij} |\widetilde{f}(i) - \widetilde{f}(j)|^2}{2\ \mathrm{Tr}\ \mathcal{L}}.$$

This is nothing but the graph analogue of the Von Neumann estimator of von Neumann (1941) explaining the terminology Graph Von Neumann estimator (GVN).

A second natural choice is given by $g(x) = \psi_J(x)$ corresponding to the filter at the finest scale. The resulting estimator is called High Pass Filter Von Neumann (HPFVN). The value of the estimator is easily computed from the coefficients as follows:

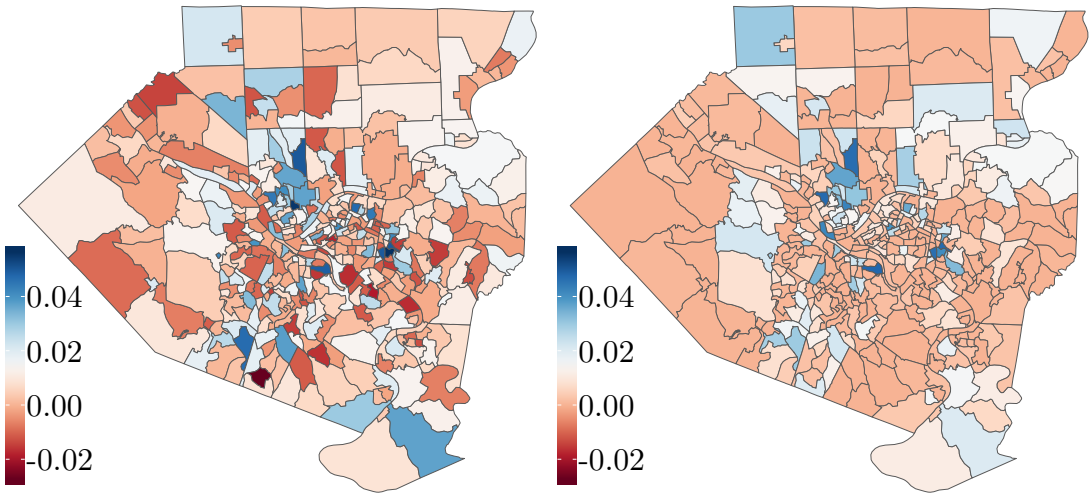
$$\hat{\sigma}_2^2 = \frac{\sum_{i=nJ+1}^{n(J+1)} (\mathcal{W}\widetilde{f})_i^2}{\mathrm{Tr}\ \psi_J(\mathcal{L})}.$$

Table 2: Mean SNR performance over $N = 10$ realizations with $f_{0.01,5}$ for the Pittsburgh graph.

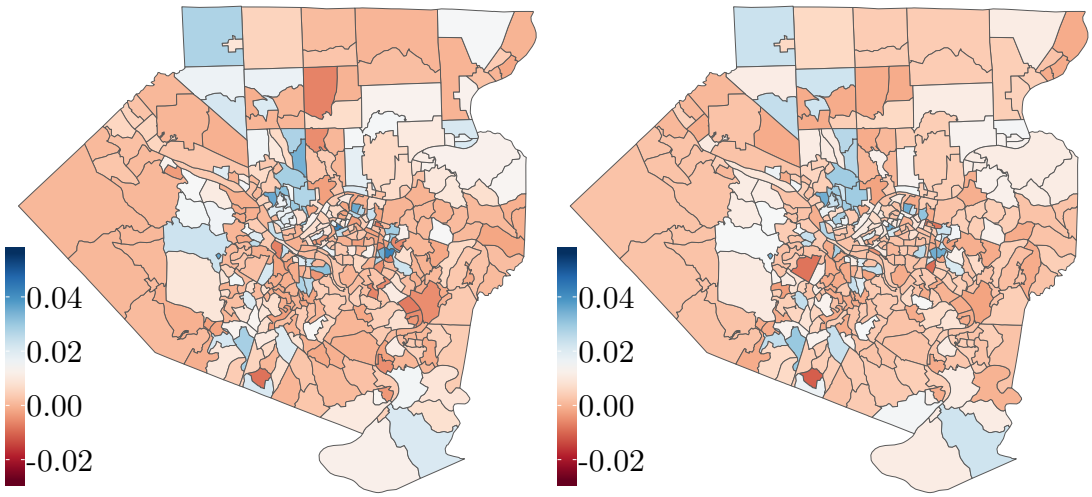
σ	0.004	0.005	0.01
$\hat{\sigma}_1$	0.0065	0.0072	0.0113
$\hat{\sigma}_2$	0.0068	0.0074	0.0114
$\mathrm{SNR}_{\mathrm{in}}$	9.53±0.29	7.60±0.29	1.58±0.29
$\mathrm{MSE}^{\beta=2,\mathrm{LD}}$	13.65±0.29	12.22±0.34	8.46±0.41
$\mathrm{MSE}^{k=2}$	12.28±0.25	11.05±0.22	8.14±0.29
$\mathrm{MSE}^{k=1}$	13.10±0.21	11.78±0.22	8.09±0.36
$\mathrm{MSE}^{k=0}$	12.83±0.23	11.42±0.24	7.56±0.46
$\mathrm{MSE}^{\mathrm{Soft}}$	11.22±0.23	9.62±0.27	5.22±0.22
$\mathrm{MSE}^{\mathrm{Hard}}$	10.22±0.28	8.50±0.38	4.03±0.14
$\mathrm{SURE}_{\sigma}^{\beta=2,\mathrm{LD}}$	13.33±0.37	12.00±0.30	7.95±0.25
$\mathrm{SURE}_{\hat{\sigma}_1}^{\beta=2,\mathrm{LD}}$	12.35±0.53	11.38±0.62	8.13±0.35
$\mathrm{SURE}_{\hat{\sigma}_2}^{\beta=2,\mathrm{LD}}$	12.12±0.57	11.27±0.54	8.07±0.33

Table 2 summarizes the findings with the nomenclature of Table 1 (where LD stands for level-dependent (LD) coordinatewise thresholding) of the main document. Lines $\mathrm{SURE}_{\hat{\sigma}_{1,2}}^{\beta=1,2}$ stand for the SURE procedure in which the noise level is estimated by the GVN ($\hat{\sigma}_1$) and the HPFVN ($\hat{\sigma}_2$). HPFVN and GVN provide a very comparable sigma estimate in this setting. Additionally, a visual comparison of our methodology with the fused lasso is illustrated in Figure 4 (corresponding to one realization in the context of the third column of the Table 2). We can see that our approach provides a gain of about 2.5dB compared to the fused lasso.

Again, in these experiments, the multiscale analysis shows better performances than the trend filtering. Besides, the estimation of σ is sufficiently accurate to have a fully data-driven procedure.



(a) $\tilde{f}$: $\mathrm{SNR}_{\mathrm{in}} = 1.97\mathrm{dB}$ (b) $f_{0.01,5}$



(c) $\mathrm{MSE}^{\beta=2}$: 9.75dB (d) $\mathrm{MSE}^{k=0}$: 7.17dB

Figure 4: Typical reconstruction for the Allegheny County example.

To conclude, let us stress that the SGWT is computationally more efficient than the GTF ($k = 1, 2$). For the latter, on a standard laptop (Intel Core i7@2.7GHz-16Go LP-DDR3@2133MHz), each 10 realizations consumed about 4h16m cpu time in mean (min:1h45m, max:6h19m) with the same grid search as for the Pittsburg in Wang et al. (2016). The CPU consumption for $k = 0$ is more decent with a mean of 3m for each 10 realizations exploiting the idea and the c++ code of Chambolle and Darbon (2009). Incidentally, this main drawback of GTF was noticed in Padilla et al. (2018) forcing a preprocessing of the graph using DFS algorithm. The SGWT on its side consumed 3m for the diagonalization and 42s for each 10 realizations.

4.4 Real Dataset: New York City Taxis

Our methodology has been also tested on a real data fetched from NYC taxis[5] databases. We build a graph with 265 vertices consisting of the LocationID (Pick-Up and Drop-Off) and define Gaussian weights $w_{ij} = \exp(-\tau d_{i,j}^2)$ where $d_{i,j}$ is the mean distance taken on all the trips between i and j or j and i. The signal f considered is defined upon the variable "total amount" on which an artificial noise is added. For an average input $\mathrm{SNR}_{\mathrm{in}} = 5.23 \pm 0.38\mathrm{dB}$ on $N = 25$

[5] https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2018-01.csv

observations, we obtain for the level dependent SURE with $\beta = 2$ and σ known, an output SNR of 10.41 ± 0.62dB compared to the performance of the oracle Wiener filter, SNR $= 7.50 \pm 0.40$dB and the oracle fused lasso, SNR $= 5.69 \pm 0.43$dB.

4.5 Correlated Noise

On the Minnesota roads graph and for the signal $f_{\eta,k}$ of Section 4 with $\eta = 0.01$ and $k = 2$, we add one realization of a correlated noise with covariance matrix given by (3) where $\alpha = 0.5$ is added leading to a noisy signal with $\text{SNR}_{\text{in}} = 2.07$dB. We run the level-dependent coordinate-wise thresholding process. The SNR given by the oracle involving the unknown signal $f_{\eta,k}$ and the one given by the SURE estimator adapted to correlated noise are similar: 8.54dB *versus* 8.51dB. In this case, the SURE for uncorrelated noise shows very bad performances: we found 4.33 for the corresponding SNR.

The quality of the SURE for uncorrelated noise is closely related to the intensity of the correlation tuned by the parameter α. As an example, if we choose $\alpha = 0.1$, the SURE adapted to correlated noise still performs very well with an estimated SNR of 10.44dB compared to the oracle 10.46dB. The SURE for uncorrelated noise is nonetheless not that bad since it estimates the SNR at 9.78dB.

Consequently, the SURE estimate for uncorrelated noise is robust to small correlations which is particularly interesting in applications since it can be difficult to estimate the correlation structure.

4.6 Further Experiments with Block Thresholding

Finally, we report some experimental results in the context of block thresholding in the setting of Section 4.3. For each scale $j = 0, 1, \ldots, 7$, the n wavelet coefficients are split into L blocks of uniform length (except for the last block that can be shorter). The best performance of block thresholding is achieved for blocks of size $|L| = 47$. The $\text{SNR}_{\text{in}} = 9.64 \pm 0.33$dB for 25 realizations. For the oracle global coordinate-wise threshold with $\beta = 2$, we obtain a SNR of 11.07 ± 0.33dB compared to the block procedure with a uniform threshold 11.82 ± 0.40dB. The block procedure performs better than the coordinate-wise one for a uniform threshold but is actually worse compared to the level dependent coordinate-wise thresholding process. The level-dependent method is expected to give better performance, but would require a more sophisticated optimization algorithm than grid search to be computationally acceptable.

5 Conclusion and Perspectives

In this paper, we have introduced a version of the SURE designed for SGWT, allowing automatic parameter selection in denoising tasks of signals on graphs. Closed-form expressions for coordinatewise and block SUREs have been provided for a wide range of threshold rules. Finally, the case of a correlated noise in the graph has also been considered. Many experiments on the Minnesota graph, the Facebook graph built, the Pittsburgh graph and the NYC taxis graph from real data are conducted.

For signals of different regularity on those graphs, it has been shown that the SURE provides an efficient estimate of the theoretical MSE. These experiments also show that multi-scale analysis is a serious competitor to existing methods. Indeed, the SGWT shows performances equivalent and sometimes even much better than the GTF, especially at high noise levels.

To be complete, the variance parameter σ^2 should be estimated. This has been (partially) addressed in the supplementary material by introducing the two estimators HPFVN and GVN. The performance of these estimators highly depends on the underlying signal. Further investigations seem to be necessary.

Theoretically, the coordinatewise soft thresholding in the transformed domain satisfies an oracle inequality as shown in Göbel et al. (2018). Similarly to the regular case, this oracle inequality states that the estimator mimics the oracle projection up to a log factor. The proof relies on the fact that the multivariate risk is expressed as a sum of univariate risks so that the Donoho's machinery applies. Using this fact, a maximal inequality for the SURE might be stated as well.

Regarding, the numerical complexity, the main limitation is the need of a complete reduction of the Laplacian. In the same vein as Hammond et al. (2011), many of the involved steps might be numerically optimized using Chebyshev polynomials. Actually, the only problematic step in the method is the computation of the weights $(\mathcal{W}\mathcal{W}^*)_{i,j}$ appearing in the SURE. However, their expression in terms of covariance suggests that Monte-Carlo estimation could work. Besides, the space-time complexity might be reduced taking advantage of the low-rank property of $\mathcal{W}\mathcal{W}^*$ implying several linear constraints on the weights (precisely nJ). Finally, for the block procedure to be completely useful, an adapted optimization algorithm should be implemented.

Some questions not addressed in the paper remains open. As already announced in introduction of the paper, the choice of a suitable frame for different graphs and different families of signals is still an open problem in spite of advances in recent years. Most likely, good choices of frame should involve a notion of graph limit such as the one introduced for graphons (see Lovász (2012)) or the more probabilistic Benjamin-Schramm limit introduced in Benjamini and Schramm (2001). This formal study should also give rise to less nave estimators of the noise level and above all give recommendations according to the class of signals considered.

References

Behjat, H., Richter, U., Van De Ville, D., and Sörnmo, L. (2016). Signal-adapted tight frames on graphs. *IEEE Trans. Signal Process.*, 64(22):6017–6029.

Belkin, M. and Niyogi, P. (2008). Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. System Sci.*, 74(8):1289–1308.

Benjamini, I. and Schramm, O. (2001). Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:no. 23, 13.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2014). Spectral networks and locally connected networks on graphs. In *ICLR*.

Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924.

Chambolle, A. and Darbon, J. (2009). On total variation minimization and surface evolution using parametric maximum flows. *International journal of computer vision*, 84(3):288.

Chung, F. R. K. (1997). *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30.

de Loynes, B., Navarro, F., and Olivier, B. (2020). Gasper: Graph signal processing in r. *arXiv preprint arXiv:2007.10642*.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, pages 3844–3852.

Deutsch, S., Ortega, A., and Medioni, G. (2016). Manifold denoising based on spectral graph wavelets. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4673–4677. IEEE.

Dong, B., Jiang, Q., Liu, C., and Shen, Z. (2016). Multiscale representation of surfaces by tight wavelet frames with applications to denoising. *Appl. Comput. Harmon. Anal.*, 41(2):561–589.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224.

Gavish, M., Nadler, B., and Coifman, R. R. (2010). Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *ICML*, pages 367–374.

Göbel, F., Blanchard, G., and von Luxburg, U. (2018). Construction of tight frames on graphs and application to denoising. In *Handbook of big data analytics*, Springer Handb. Comput. Stat., pages 503–522. Springer, Cham.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034.

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.*, 30(2):129–150.

Henaff, M., Bruna, J., and LeCun, Y. (2015). Deep convolutional networks on graph-structured data. In *NIPS*.

Irion, J. and Saito, N. (2017). Efficient approximation and denoising of graph signals using the multiscale basis dictionaries. *IEEE Trans. Signal Inform. Process. Netw.*, 3(3):607–616.

Kereta, Z., Vigogna, S., Naumova, V., Rosasco, L., and De Vito, E. (2019). Monte carlo wavelets: a randomized approach to frame discretization. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.

Leonardi, N. and Van De Ville, D. (2013). Tight wavelet frames on multislice graphs. *IEEE Trans. Signal Process.*, 61(13):3357–3367.

Leskovec, J. and Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547.

Liu, J. S. (1994). Siegel's formula via Stein's identities. *Statist. Probab. Lett.*, 21(3):247–251.

Lovász, L. (2012). *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI.

Luisier, F., Blu, T., and Unser, M. (2007). A new SURE approach to image denoising: interscale orthonormal wavelet thresholding. *IEEE Trans. Image Process.*, 16(3):593–606.

Mallat, S. (2009). *A wavelet tour of signal processing*. Elsevier/Academic Press, Amsterdam, third edition. The sparse way, With contributions from Gabriel Peyré.

Navarro, F., Fadili, M., and Chesneau, C. (2013). Adaptive parameter selection for block wavelet-thresholding deconvolution. *IFAC Proceedings Volumes*, 46(11):495–499.

Onuki, M., Ono, S., Yamagishi, M., and Tanaka, Y. (2016). Graph signal denoising via trilateral filter on graph spectral domain. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):137–148.

Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., and Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828.

Padilla, O. H. M., Sharpnack, J., Scott, J. G., and Tibshirani, R. J. (2018). The dfs fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:1–36.

Pesquet, J.-C., Benazza-Benyahia, A., and Chaux, C. (2009). A SURE approach for digital signal/image deconvolution problems. *IEEE Trans. Signal Process.*, 57(12):4616–4632.

Peyré, G., Fadili, J., and Chesneau, C. (2011). Adaptive structured block sparsity via dyadic partitioning. In *2011 19th European Signal Processing Conference*, pages 1455–1459. IEEE.

Sharpnack, J., Singh, A., and Krishnamurthy, A. (2013). Detecting activations over graphs using spanning tree wavelet bases. In *Artificial Intelligence and Statistics (AISTATS-13)*, pages 536–544.

Shuman, D., Narang, S., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 3(30):83–98.

Shuman, D. I., Wiesmeyr, C., Holighaus, N., and Vandergheynst, P. (2015). Spectrum-adapted tight graph wavelet and vertex-frequency frames. *IEEE Transactions on Signal Processing*, 63(16):4223–4235.

Smola, A. J. and Kondor, R. (2003). Kernels and regularization on graphs. In *Learning theory and kernel machines*, pages 144–158. Springer.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.*, 39(3):1335–1371.

Tremblay, N. and Borgnat, P. (2014). Graph wavelets for multiscale community mining. *IEEE Trans. Signal Process.*, 62(20):5227–5239.

Vaiter, S., Deledalle, C.-A., Peyré, G., Dossal, C., and Fadili, J. (2013). Local behavior of sparse analysis regularization: applications to risk estimation. *Appl. Comput. Harmon. Anal.*, 35(3):433–451.

von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statistics*, 12:367–395.

Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.*, 17:Paper No. 105, 41.