Denoising by Higher Order Statistics

Tanja Teuber[a], Steffen Remmele[b], Jürgen Hesser[b], Gabriele Steidl[a]

[a] *Mathematical Image Processing and Data Analysis Group, University of Kaiserslautern, Germany*
[b] *Medical Centre Mannheim, University of Heidelberg, Mannheim, Germany*

Abstract

A standard approach for deducing a variational denoising method is the maximum a posteriori strategy. Here, the denoising result is chosen in such a way that it maximizes the conditional density function of the reconstruction given its observed noisy version. Unfortunately, this approach does not imply that the empirical distribution of the reconstructed noise components follows the statistics of the assumed noise model. In this paper, we propose to overcome this drawback by applying an additional transformation to the random vector modeling the noise. This transformation is then incorporated into the standard denoising approach and leads to a more sophisticated data fidelity term, which forces the removed noise components to have the desired statistical properties. The good properties of our new approach are demonstrated for additive Gaussian noise by numerical examples. Our method shows to be especially well suited for data containing high frequency structures, where other denoising methods which assume a certain smoothness of the signal cannot restore the small structures.

Keywords:
denoising, additive Gaussian noise, maximum a posteriori estimation, higher-order moments

1. Introduction

Measured signals and images are usually corrupted by noise which makes their denoising and reconstruction a central aim in signal and image processing. Especially data with a low quality requires reliable and robust reconstruction methods. In the last decades many methods have been proposed for denoising corrupted data. A common approach is to solve a variational

problem, where one has to minimize a functional consisting of a data fidelity term and a regularization term. The functional is usually deduced by a maximum a posteriori strategy, which requires some knowledge about the noise statistics and the distribution of the original data. In literature, e.g., when considering detector noise or in case of high photon counts, where the Poisson distribution can be well approximated by a Gaussian one, it is often assumed that the corrupted data follows an additive Gaussian noise model. This means that our given noisy data $g \in \mathbb{R}^N$ can be modeled as

$$g = f_0 + \varepsilon_0,$$

where $f_0 \in \mathbb{R}^N$ is the unknown noise-free data and $\varepsilon_0 \in \mathbb{R}^N$ is the realization of a random vector $\mathcal{E} : \Omega \to \mathbb{R}^N$ defined with respect to a continuous probability space $(\Omega, \mathcal{F}, P)$. Here, Ω denotes a sample space, $\mathcal{F}$ a σ-algebra and $P : \mathcal{F} \to [0, 1]$ a probability measure. The vectors g and f_0 are assumed to be realizations of independent N-dimensional random vectors $G : \Omega \to \mathbb{R}^N$ and $F : \Omega \to \mathbb{R}^N$, respectively, so that $G = F + \mathcal{E}$.

To deduce an estimate $\hat{f}_{MAP}$ of f_0 by a *maximum a posteriori (MAP)* strategy, one usually sets

$$\hat{f}_{MAP} \;:=\; \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \{ -\log p_{F|G}(f|g) \}, \tag{1}$$

where $p_{F|G}(f|g)$ is the conditional probability density for observing f given $G = g$. By Bayes' theorem it holds that

$$p_{F|G}(f|g) \;=\; \frac{p_{G|F}(g|f)\, p_F(f)}{p_G(g)}. \tag{2}$$

Here, $p_{G|F}$ is the so-called *likelihood*, which is usually closely related to the density of the noise, p_F is some a priori density of F and p_G is the density of G. Since we consider additive noise, it holds that $p_{G|F}(g|f) = p_{\mathcal{E}}(g - f) = p_{\mathcal{E}}(\varepsilon)$, where $\varepsilon := g - f$ and $p_{\mathcal{E}}$ denotes the density of $\mathcal{E}$. Moreover, by inserting (2) in (1) it follows that

$$\hat{f}_{MAP} \;=\; \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \{ -\log p_{\mathcal{E}}(g - f) - \log p_F(f) \}. \tag{3}$$

Here, the terms $-\log p_{\mathcal{E}}(g - f)$ and $-\log p_F(f)$ imply that we search for the most likely vectors $\hat{\varepsilon}_{MAP} = g - \hat{f}_{MAP}$ and $\hat{f}_{MAP}$ under the condition that

$g = \hat{f}_{MAP} + \hat{\varepsilon}_{MAP}$. If the components $\mathcal{E}_i$ of the random vector $\mathcal{E}$ are pairwise independent and identically distributed (i.i.d.) as it is often assumed, then

$$-\log p_{\mathcal{E}}(g-f) \;=\; -\log \prod_{i=1}^{N} p_{\mathcal{E}_i}(g_i - f_i) \;=\; -\sum_{i=1}^{N} \log p_{\mathcal{E}_i}(g_i - f_i). \qquad (4)$$

For the special case that $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, N$, this leads to

$$\begin{aligned} -\log p_{\mathcal{E}}(g-f) &= -\sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(g_i - f_i)^2}{2\sigma^2} \right) \\ &= -N \log \frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2} \|g - f\|_2^2. \end{aligned} \qquad (5)$$

To determine $-\log p_F(f)$, at least some estimate of the a priori density p_F is needed. Assuming that $p_F(f) = \exp(-c\,J(f))$ for some constant $c > 0$ and a nonnegative function $J : \mathbb{R}^N \to \mathbb{R}$, the minimization problem (3) with (5) is finally equivalent to

$$\hat{f}_{\mathrm{MAP}} \;=\; \operatorname*{argmin}_{f \in \mathbb{R}^N} \{ \frac{1}{2} \|g - f\|_2^2 + \lambda\, J(f) \} \qquad \text{with} \quad \lambda := c\,\sigma^2 > 0. \qquad (6)$$

Here, the amount of filtering is controlled by the parameter λ, which steers the influence of the two terms within the functional. If J is assumed to be $J(f) := \|Df\|_2^2$, where D is a discrete first derivative operator, we obtain by this approach the regularization method proposed by Tikhonov and Miller (TM) in [1], which we will shortly call MAP-TM. By this choice for J the initial signal is assumed to have small first derivatives, i.e., to be of a certain degree of smoothness (in H^1 for the continuous setting). Unfortunately, if the signal contains jumps, the TM regularization will oversmooth them. To overcome this drawback, J is often set to $J(f) := \|Df\|_1$, which is the discrete one-dimensional version of the total variation regularizer (TV). The corresponding denoising method (6) leads to the classical approach of Rudin, Osher and Fatemi [2], which is well known for its discontinuity preserving properties. In the following, we will refer to this method as MAP-TV and we will use it as well as the MAP-TM approach as reference methods for our numerical experiments.

Now, if we forget about the regularization term for a moment and have again a closer look at our data fidelity term $-\log p_{\mathcal{E}}(g-f)$ in (4), where

$\mathcal{E}$ is assumed to be i.i.d., we see that this data fidelity term is minimal whenever all components $\varepsilon_i = g_i - f_i$ maximize $p_{\mathcal{E}_i}(\varepsilon_i)$. Consequently, without the regularization term or equivalently for $\lambda = 0$, our reconstructed noise vector $\hat{\varepsilon}$ would be a constant vector of value $\operatorname{argmax}_e p_{\mathcal{E}_i}(e)$ and thus, $\hat{f} = g - \operatorname{argmax}_e p_{\mathcal{E}_i}(e)$. These estimates may seem reasonable for a signal length N close to one. However, since the vector $\mathcal{E}$ is i.i.d., we may expect for larger N that the empirical distribution of the components of our estimated noise vector $\hat{\varepsilon}$ resembles the distribution of $\mathcal{E}_i$, $i = 1, \ldots, N$. In principal, to check how good a set of samples coincides with a given distribution we could for example apply the Kolmogorov-Smirnov [3] or the Anderson-Darling test [4].

In contrast to [5, 6] we introduce a representation of the noise distribution that depends both on moments and especially on the correlation of the random variables $\mathcal{E}_i$. We also embed noise correlation [7] in a concise formalism that allows to achieve de-correlated estimates ε_i of the original noise components if the $\mathcal{E}_i$ are independent, a result that is often achieved ad hoc with non-local means [8] according to empirical studies.

Outline. In the following, we show that it is possible to overcome the drawbacks of the standard MAP approach by applying a suitable variable transformation to the random vector $\mathcal{E}$ before computing the MAP estimates. Our new approach is presented in Section 2 and we investigate two different transformations with respect to their benefits and shortcomings. In Section 3 we discuss a first implementation of our approach for one-dimensional data and present numerical results. Finally, we summarize our new findings and finish with concluding remarks in Section 4.

2. A new denoising approach

For simplicity, we assume in the following that the random variables $\mathcal{E}_i$ are again i.i.d. with expectation value $E(\mathcal{E}_i) = \mu$ and variance $\operatorname{Var}(\mathcal{E}_i) = \sigma^2$. Hence, the components of the vector ε can be considered as samples of the same random variable. Computing the MAP estimator $\hat{f}_{MAP}$ and the corresponding noise vector $\hat{\varepsilon}_{MAP} = g - \hat{f}_{MAP}$ from equation (3) is equivalent to solving the minimization problem

$$\underset{f \in \mathbb{R}^N,\, \varepsilon \in \mathbb{R}^N}{\operatorname{argmin}} \{-\log p_{\mathcal{E}}(\varepsilon) - \log p_F(f)\} \qquad \text{subject to} \quad g = f + \varepsilon \tag{7}$$

for the given noisy data $g \in \mathbb{R}^N$. Since the term $-\log p_{\mathcal{E}}(\varepsilon)$ does not guarantee that the empirical distribution of the components of the estimate $\hat{\varepsilon}_{MAP}$

approximates the distribution of $\mathcal{E}_i$, $i = 1, \ldots, N$, as demonstrated above, we want to replace it by a term with better properties. One approach in this direction can be found for example in [9], where the authors propose to incorporate a multiresolution statistic into the data fidelity term. The basic idea for our approach is to transform the random vector $\mathcal{E}$ into a new random vector $V := T(\mathcal{E})$ with corresponding density p_V using an appropriate transformation $T : \mathbb{R}^N \to \mathbb{R}^M$ and to solve instead of (7) the minimization problem

$$\underset{f \in \mathbb{R}^N,\, \varepsilon \in \mathbb{R}^N}{\operatorname{argmin}} \; \{-\log p_V(T(\varepsilon)) - \log p_F(f)\} \qquad \text{subject to} \;\; g = f + \varepsilon. \tag{8}$$

Obviously, if $M = N$ and T is the identity, then $V = \mathcal{E}$ and problems (7) and (8) coincide. A more general result is given by the following proposition.

Proposition 1. *If T is injective and has a continuously differentiable inverse T^{-1} on its range with non-vanishing Jacobian $J_{T^{-1}}$, then*

$$p_V(T(\varepsilon)) \;=\; p_{\mathcal{E}}(\varepsilon)\, |\det J_{T^{-1}}(T(\varepsilon))|.$$

Thus, if additionally $|\det J_{T^{-1}}(T(\varepsilon))| = 1$ for all $\varepsilon \in \mathbb{R}^N$, then problems (7) and (8) are equivalent.

This proposition follows directly from Jacobi's transformation formula, see e.g. [10, p. 92f]. Of course, the interesting cases are those, where (7) and (8) are not equivalent.

Remark 1. *Although we have assumed that the random variables $\mathcal{E}_i$ are i.i.d., in the often considered case of normally distributed random variables, this restriction, in particular the independence, can be omitted. In fact, we can exploit that whenever $\mathcal{E} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with mean vector $\boldsymbol{\mu} \in \mathbb{R}^N$ and covariance matrix $\Sigma \in \mathbb{R}^{N,N}$, then the components $\mathcal{E}_i$ of $\mathcal{E}$ are pairwise independent if and only if Σ is a diagonal matrix, see, e.g., [11, p. 214]. Since Σ is symmetric, there exists further an eigenvalue decomposition of Σ such that $\Sigma = U D^{\frac{1}{2}} D^{\frac{1}{2}} U^T$ for some orthogonal matrix $U \in \mathbb{R}^{N,N}$ and a diagonal matrix $D^{\frac{1}{2}} \in \mathbb{R}^{N,N}$. By setting $\widetilde{\mathcal{E}} := T_0(\mathcal{E})$ with $T_0(\mathcal{E}) = D^{-\frac{1}{2}} U^T (\mathcal{E} - \mu)$, where $D^{-\frac{1}{2}}$ denotes the pseudoinverse of $D^{\frac{1}{2}}$ if the inverse does not exist, one can show that $\widetilde{\mathcal{E}} \sim \mathcal{N}(0_N, I_N)$ is i.i.d. with 0_N denoting a vector consisting of N zeros and I_N being the identity matrix of size $N \times N$. Hence, we only have to replace the transformation T used in (8) by $\widetilde{T} = T \circ T_0$ to handle also random vectors $\mathcal{E}$ which are normally distributed, but where the components are not i.i.d.*

In the following, we propose two transformations which are useful for our purposes. The first transform is theoretically well-suited, but the numerical realization is only feasible for huge datasets. The second transform can be considered as a modification of the first one which is computationally more capable.

Moment transformation. Our first transformation T is given by

$$(v_1, \dots, v_M)^{\mathrm{T}} = T(\varepsilon_1, \dots, \varepsilon_N) := \left(\tfrac{1}{N} \sum_{i=1}^{N} \varepsilon_i,\ \tfrac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2, \dots, \tfrac{1}{N} \sum_{i=1}^{N} \varepsilon_i^M \right)^{\mathrm{T}}, \quad (9)$$

i.e. v_k is an estimate of the expectation value $E(\mathcal{E}_i^k)$, the kth raw moment of the $\mathcal{E}_i$, $i = 1, \dots, N$.

Remark 2. *It is well known that alternatively to characterizing $\mathcal{E}_i$ by its density function, it is also uniquely determined by its moment generating function $\mathcal{M}_{\mathcal{E}_i}(t) := E(e^{[t\mathcal{E}_i]})$ if $\mathcal{M}_{\mathcal{E}_i}$ is finite in some open ball around zero, see, e.g., [11]. In this case, the kth moment $m_{\mathcal{E}_i}(k) := E(\mathcal{E}_i^k)$ of $\mathcal{E}_i$ can be deduced from its moment generating function by $m_{\mathcal{E}_i}(k) = \frac{\partial^k}{\partial t^k} \mathcal{M}_{\mathcal{E}_i}(0)$ and the moment generating function is itself uniquely determined by its moments via the Taylor series expansion. In detail, we have*

$$\mathcal{M}_{\mathcal{E}_i}(t) \ = \ \sum_{k=0}^{\infty} \frac{t^k}{k!}\, m_{\mathcal{E}_i}(k) \ = \ \sum_{k=0}^{\infty} \frac{t^k}{k!}\, E(\mathcal{E}_i^k).$$

If for example $\mathcal{E}_i \sim \mathcal{N}(\mu, \sigma^2)$ is normally distributed with mean value μ and variance σ^2, the moment generating function exists and it is given by

$$\mathcal{M}_{\mathcal{E}_i}(t) \ = \ \exp(t\mu + \tfrac{1}{2}\sigma^2 t^2) \qquad \forall\, t \in \mathbb{R}.$$

By this remark we see that for a sufficiently large number of moments M the transformation (9) preserves, loosely speaking, the information contained in $(\varepsilon_1, \dots, \varepsilon_N)^{\mathrm{T}}$ about the density $p_{\mathcal{E}}$, although T is not even injective. Note that any of the $N!$ possible permutations of the values $\varepsilon_1, \dots, \varepsilon_N$ within the vector ε leads to the same value $T(\varepsilon)$. However, this implies that whenever two vectors ε, $\widetilde{\varepsilon}$ have the same empirical distribution, then $p_V(T(\varepsilon)) = p_V(T(\widetilde{\varepsilon}))$. Moreover, $p_V(T(\varepsilon))$ is supposed to be large whenever the estimated moments $T(\varepsilon)$ are close to the moments of the $\mathcal{E}_i$, $i = 1, \dots, N$. For our data fidelity term $-\log p_V(T(\varepsilon))$ this implies that, loosely speaking,

it does not favor any particular noise vector as long as the empirical distribution of ε approximates well the distribution of $\mathcal{E}_i$. Hence, the role of $-\log p_F(f)$ in (8) is to choose a particular noise vector such that $p_F(f)$ is large for the resulting data vector f.

Unfortunately, the minimization problem (8) with the particular transformation (9) has the following drawback: To obtain a reliable estimate $v_k = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^k$ of the kth moment $E(\mathcal{E}_i^k)$, the number of necessary samples N can significantly increase with k as we will see in the following. To this end, we determine the variance of the random variables $V_k := \frac{1}{N}\sum_{i=1}^{N} \mathcal{E}_i^k$.

Lemma 1. *Let X_i, $i = 1, \ldots, N$, be i.i.d. random variables with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then, for $M_k := \frac{1}{N}\sum_{i=1}^{N} X_i^k$, $k \in \mathbb{N}$ and $i = 1, \ldots, N$ it holds that*

$$E(M_k) \;=\; E(X_i^k) \quad \text{and} \quad \mathrm{Var}(M_k) \;=\; \frac{1}{N}(E(X_i^{2k}) - E(X_i^k)^2).$$

In particular, for $k = 1$ we have

$$E(M_1) \;=\; \mu \quad \text{and} \quad \mathrm{Var}(M_1) \;=\; \frac{1}{N}(E(X_i^2) - E(X_i)^2) \;=\; \frac{\sigma^2}{N}.$$

The proof of this lemma follows directly by applying standard results for the calculation with expectation values and variances.

If we assume for example that $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$, then

$$E(\mathcal{E}_i^k) \;=\; \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 1 \cdot 3 \cdots (k-1)\,\sigma^k & \text{if } k \text{ is even,} \end{cases} \tag{10}$$

see, e.g., [11, p. 93]. As a consequence, we obtain

$$\mathrm{Var}(V_1) = \frac{\sigma^2}{N}, \quad \mathrm{Var}(V_2) = \frac{2\sigma^4}{N}, \quad \mathrm{Var}(V_3) = \frac{15\sigma^6}{N}, \quad \mathrm{Var}(V_4) = \frac{96\sigma^8}{N}, \ldots$$

For $\sigma^2 \geq 1$ and a constant number of samples N, this implies that the variance increases significantly with the order k of the estimated moments. Thus, to keep the variances constant, the number of samples has to increase significantly with k. Already for $\sigma^2 = 1$ and N samples used for estimating V_1, we would need $2N$ samples for V_2, $15N$ samples for V_3 and $96N$ samples for estimating V_4.

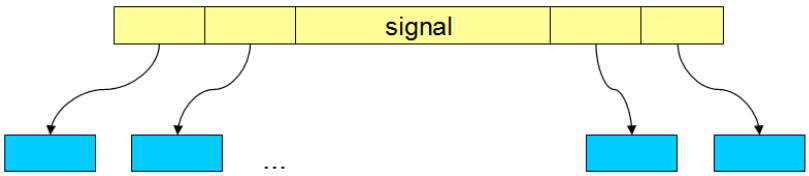

Figure 1: Decomposition of ε into equally sized subvectors $\tilde{\varepsilon}_k$ for $k = 1, \ldots, K$.

Modified moment transformation. To overcome this drawback of transformation (9) while still exploiting its benefits, we consider a different transformation. In the following, we want to restrict our attention to estimates of the first two moments. For $N = K\widetilde{N}$ with $K, \widetilde{N} \in \mathbb{N}$, we split the random vector $\mathcal{E}$ into K equally sized subvectors of length $\widetilde{N}$ as

$$\mathcal{E} := \left(\tilde{\mathcal{E}}_k\right)_{k=1}^{K} \quad \text{with} \quad \tilde{\mathcal{E}}_k := \left(\mathcal{E}_{(k-1)\widetilde{N}+j}\right)_{j=1}^{\widetilde{N}}$$

and similarly the vector of realizations $\varepsilon = (\tilde{\varepsilon}_k)_{k=1}^{K}$, cf. Figure 1. Let $\mu := E(\mathcal{E}_i)$ and $\mu_k := \frac{1}{\widetilde{N}} \sum_{j=1}^{\widetilde{N}} \varepsilon_{(k-1)\widetilde{N}+j}$. Then, we define our transform $T = (T_k)_{k=1}^{K} : \mathbb{R}^N \to \mathbb{R}^{\frac{K(K+3)}{2}}$ by

$$T_k(\varepsilon) := \begin{pmatrix} \mu_k \\ \frac{1}{\widetilde{N}} \left(\langle \tilde{\varepsilon}_k - \mu, \tilde{\varepsilon}_l - \mu \rangle\right)_{l=k}^{K} \end{pmatrix}.$$

The transformation T_k maps ε to estimates of the mean and variance of the $\tilde{\mathcal{E}}_k$ as well as to estimates of the covariances between the random variables of the vectors $\tilde{\mathcal{E}}_k$ and $\tilde{\mathcal{E}}_l$ for $l > k$. Since the vectors $\tilde{\mathcal{E}}_j$ are supposed to be independent, for $\widetilde{N}$ large enough the matrix $\frac{1}{\widetilde{N}}(\langle \tilde{\varepsilon}_k - \mu, \tilde{\varepsilon}_l - \mu \rangle)_{k,l=1}^{K}$ is approximately a diagonal matrix with a nearly constant diagonal.
Of course other splittings than just those into subsequent vectors as well as multiple splittings could also be used in the construction of T, cf. Section 4.

In order to incorporate T into (8) we need to determine the density p_V of the transformed random vector $V = T(\mathcal{E})$. To this purpose, we use the notation $T_k(\mathcal{E}) = (M_k, C_{k,k}, C_{k,k+1}, \ldots, C_{k,K})$.

For a Gaussian distributed random vector $\mathcal{E} \sim \mathcal{N}(\mu, \sigma^2)$ it holds that $M_k \sim \mathcal{N}(\mu, \frac{1}{\widetilde{N}}\sigma^2)$ and $\widetilde{N}/\sigma^2 \cdot C_{k,k}$ is χ^2-distributed with $\widetilde{N}$ degrees of freedom. Moreover, M_k and $C_{k,k}$ are independent if we replace μ by the sample mean M_k in the definition of $C_{k,k}$, cf. [11, Thm. 4.4.2].

However, in general we cannot expect the components of V to be mutually independent and thus, the density p_V is not just the product of the densities of the random variables forming the vector V. Therefore, we use the following proposition to determine an estimate of p_V:

Proposition 2. *Let X_i, Y_i, Z_i, $i = 1, \ldots, N$, be i.i.d. random variables with expectation value μ and variance σ^2. Moreover, set*

$$\begin{pmatrix} M \\ S \\ C_{XY} \\ C_{XZ} \end{pmatrix} := \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} X_i \\ (X_i - \mu)^2 \\ (X_i - \mu)(Y_i - \mu) \\ (X_i - \mu)(Z_i - \mu) \end{pmatrix}.$$

Then, the random vector $(M, S, C_{XY}, C_{XZ})^T$ has mean vector $\boldsymbol{\mu} = (\mu, \sigma^2, 0, 0)^T$ and covariance matrix

$$\Sigma = \frac{1}{N} \begin{pmatrix} \sigma^2 & E\big((X_i - \mu)^3\big) & 0 & 0 \\ E\big((X_i - \mu)^3\big) & E\big((X_i - \mu)^4\big) - \sigma^4 & 0 & 0 \\ 0 & 0 & \sigma^4 & 0 \\ 0 & 0 & 0 & \sigma^4 \end{pmatrix}. \tag{11}$$

Moreover, it is asymptotically normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ. In the particular case of normally distributed random variables, the covariance matrix is given by

$$\Sigma = \frac{1}{N} \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & 2\sigma^4 & 0 & 0 \\ 0 & 0 & \sigma^4 & 0 \\ 0 & 0 & 0 & \sigma^4 \end{pmatrix}. \tag{12}$$

Proof: By Lemma 1 and since

$$E(aX + b) = aE(X) + b, \quad E(X + Y) = E(X) + E(Y)$$
$$\text{and} \quad E(XY) = E(X)E(Y) + \mathrm{Cov}(X, Y),$$

where $\mathrm{Cov}(X, Y) = 0$ in case of independent random variables with finite variance, we obtain that $\boldsymbol{\mu} = (\mu, \sigma^2, 0, 0)^{\mathrm{T}}$. By definition of the covariance matrix we have

$$\Sigma = \begin{pmatrix} \mathrm{Var}(M) & \mathrm{Cov}(M, S) & \mathrm{Cov}(M, C_{XY}) & \mathrm{Cov}(M, C_{XZ}) \\ \mathrm{Cov}(S, M) & \mathrm{Var}(S) & \mathrm{Cov}(S, C_{XY}) & \mathrm{Cov}(S, C_{XZ}) \\ \mathrm{Cov}(C_{XY}, M) & \mathrm{Cov}(C_{XY}, S) & \mathrm{Var}(C_{XY}) & \mathrm{Cov}(C_{XY}, C_{XZ}) \\ \mathrm{Cov}(C_{XZ}, M) & \mathrm{Cov}(C_{XZ}, S) & \mathrm{Cov}(C_{XZ}, C_{XY}) & \mathrm{Var}(C_{XZ}) \end{pmatrix}.$$

Now, we consider the diagonal of Σ (with similar arguments for $\mathrm{Var}(C_{XY})$ and $\mathrm{Var}(C_{XZ})$). By Lemma 1 and since

$$\mathrm{Var}(X^k) = E(X^{2k}) - E(X^k)^2 \text{ for } k \in \mathbb{N}, \quad \mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X),$$
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

and for independent X, Y further

$$\mathrm{Var}(XY) = E(Y)^2\,\mathrm{Var}(X) + E(X)^2\,\mathrm{Var}(Y) + \mathrm{Var}(X)\mathrm{Var}(Y),$$

we conclude that

$$\begin{aligned}
\begin{pmatrix} \mathrm{Var}(M) \\ \mathrm{Var}(S) \\ \mathrm{Var}(C_{XY}) \end{pmatrix} &= \begin{pmatrix} \frac{1}{N}\,\sigma^2 \\ \frac{1}{N}\Big(E\big((X_i-\mu)^4\big) - E\big((X_i-\mu)^2\big)^2\Big) \\ \frac{1}{N^2}\sum\limits_{i=1}^{N} \mathrm{Var}\Big((X_i-\mu)(Y_i-\mu)\Big) \end{pmatrix} \\
&= \frac{1}{N}\begin{pmatrix} \sigma^2 \\ E\big((X_i-\mu)^4\big) - E\big((X_i-\mu)^2\big)^2 \\ \mathrm{Var}(X_i-\mu)\,\mathrm{Var}(Y_i-\mu) \end{pmatrix} \\
&= \frac{1}{N}\begin{pmatrix} \sigma^2 \\ E\big((X_i-\mu)^4\big) - \sigma^4 \\ \sigma^4 \end{pmatrix}.
\end{aligned}$$

Furthermore, using the above rules again as well as

$$\mathrm{Cov}(a\,X, b\,Y) = ab\,\mathrm{Cov}(X,Y) \text{ and } \mathrm{Cov}\Big(\sum_{i=1}^{N} X_i, \sum_{j=1}^{N} Y_j\Big) = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Cov}(X_i, Y_j),$$

we obtain the following off-diagonal elements of Σ:

$$\begin{aligned}
\begin{pmatrix} \mathrm{Cov}(M,S) \\ \mathrm{Cov}(M,C_{XY}) \\ \mathrm{Cov}(S,C_{XY}) \\ \mathrm{Cov}(C_{XY},C_{XZ}) \end{pmatrix} &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \begin{pmatrix} \mathrm{Cov}\big(X_i,(X_j-\mu)^2\big) \\ \mathrm{Cov}\big(X_i,(X_j-\mu)(Y_j-\mu)\big) \\ \mathrm{Cov}\big((X_i-\mu)^2,(X_j-\mu)(Y_j-\mu)\big) \\ \mathrm{Cov}\big((X_i-\mu)(Y_i-\mu),(X_j-\mu)(Z_j-\mu)\big) \end{pmatrix} \\
&= \frac{1}{N^2}\sum_{i=1}^{N} \begin{pmatrix} E\big((X_i-\mu)((X_i-\mu)^2-\sigma^2)\big) \\ E\big((X_i-\mu)^2(Y_i-\mu)\big) \\ E\big(((X_i-\mu)^2-\sigma^2)(X_i-\mu)(Y_i-\mu)\big) \\ E\big((X_i-\mu)^2(Y_i-\mu)(Z_i-\mu)\big) \end{pmatrix} \\
&= \frac{1}{N} \begin{pmatrix} E\big((X_i-\mu)^3\big) - \sigma^2\,E(X_i-\mu) \\ E\big((X_i-\mu)^2\big)\,E(Y_i-\mu) \\ E\big((X_i-\mu)^3 - \sigma^2\,E(X_i-\mu)\big)\,E(Y_i-\mu) \\ E\big((X_i-\mu)^2\big)\,E(Y_i-\mu)\,E(Z_i-\mu) \end{pmatrix}
\end{aligned}$$

$$= \frac{1}{N} \begin{pmatrix} E\big((X_i - \mu)^3\big) \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Since Σ is symmetric, this leads finally to the matrix (11). The special matrix (12) follows directly from (10). Moreover, we obtain by the central limit theorem, see, e.g., [12, p. 28], that $(M, S, C_{XY}, C_{XZ})^{\mathrm{T}}$ is asymptotically normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. $\square$

In the case of normally distributed random variables, it follows directly from this result that each random vector $T_k(\mathcal{E})$, $k = 1, \ldots, K$, is uncorrelated and thus, this applies also to the whole vector $V = T(\mathcal{E})$. Besides, for N large enough we can approximate

$$\begin{aligned} p_V(T(\varepsilon)) \;\approx\; c \prod_{k=1}^{K} \exp \frac{-(\mu_k - \mu)^2}{\frac{2}{\widetilde{N}}\sigma^2} \;\exp \frac{-\left(\frac{1}{\widetilde{N}}\|\tilde{\varepsilon}_k - \mu\|^2 - \sigma^2\right)^2}{\frac{4}{\widetilde{N}}\sigma^4} \\ \cdot \prod_{k=1}^{K} \prod_{l=k+1}^{K} \exp \frac{-\left(\frac{1}{\widetilde{N}}\langle \tilde{\varepsilon}_k - \mu, \varepsilon_l - \mu\rangle\right)^2}{\frac{2}{\widetilde{N}}\sigma^4} \end{aligned}$$

with $c := (2\pi)^{K + \frac{K(K+1)}{4}}\, 2^{\frac{K}{2}}\, \widetilde{N}^{-K - \frac{K(K+1)}{4}} \sigma^{K(K+4)}$ and thus,

$$-\log p_V(T(\varepsilon)) \;\approx\; J^{(1)}_{\mathrm{mean}}(\varepsilon) + J^{(2)}_{\mathrm{var}}(\varepsilon) + J^{(2)}_{\mathrm{cov}}(\varepsilon) - \log c \tag{13}$$

with

$$\begin{aligned} J^{(1)}_{\mathrm{mean}}(\varepsilon) &= \frac{\widetilde{N}}{2\sigma^2} \sum_{k=1}^{K} (\mu_k - \mu)^2, \\ J^{(2)}_{\mathrm{var}}(\varepsilon) &= \frac{1}{4\widetilde{N}\sigma^4} \sum_{k=1}^{K} \left(\|\tilde{\varepsilon}_k - \mu\|^2 - \tilde{N}\sigma^2\right)^2, \\ J^{(2)}_{\mathrm{cov}}(\varepsilon) &= \frac{1}{2\widetilde{N}\sigma^4} \sum_{k=1}^{K} \sum_{l=k+1}^{K} \langle \tilde{\varepsilon}_k - \mu, \tilde{\varepsilon}_l - \mu\rangle^2. \end{aligned}$$

Hence, in contrast to $-\log p_V(\varepsilon)$, our new data fidelity term in (13) favor vectors, where each subvector $\tilde{\varepsilon}_k$ has approximately the mean value and variance

we would expect from the statistics. Moreover, all subvectors are forced to be uncorrelated, which forces all parts of the original signal to be contained in $\hat{f}$ rather than $\hat{\varepsilon}$ when minimizing (8). However, in contrast to (5) our new data fidelity term in (13) is no longer convex, which means that it can have many local minimal. Indeed, if ε minimizes our data fidelity term, then this is also true for all permutations of the subvectors within ε as well as for every permutation of entries within one subvector. Thus, the regularization term in (8) is again essential for restricting the possible set of solutions.

3. Numerical results

Minimizaton problem. To demonstrate the capability of our new denoising approach we proceed with numerical examples. In the following, we want to minimize the *higher order statistics (HOS) functional*

$$J(\varepsilon) \;=\; J^{(1)}_{\text{mean}}(\varepsilon) \;+\; J^{(2)}_{\text{var}}(\varepsilon) \;+\; \frac{2}{K-1}\, J^{(2)}_{\text{cov}}(\varepsilon) \;+\; \lambda \,\|g-\varepsilon\|_2^2 \qquad (14)$$

with respect to ε so that our reconstruction of the original data vector is given by $\hat{f} = g - \hat{\varepsilon}$. Here, the regularization term $\|g-\varepsilon\|_2^2$ weighted by the parameter $\lambda > 0$ guarantees that the reconstructed noise vector $\hat{\varepsilon}$ is related to the given noisy signal g. This function is kept quite simple and λ is always chosen very small to ensure that the result is mainly determined by our new higher order terms. For denoising smoother functions other smoothness term could be applied.
For our experiments we set the length of the subvectors to $\widetilde{N} = 50$ so that the number of subvectors K is given by the signal length divided by $\widetilde{N}$. Note that in (14) the functional $J^{(2)}_{cov}$ is weighted differently compared to the functional on the right-hand side of (13). Since $J^{(1)}_{mean}$ and $J^{(2)}_{var}$ are based on K summands whereas $J^{(2)}_{cov}$ consists of $\frac{K(K-1)}{2}$ values, we multiply $J^{(2)}_{cov}$ by $\frac{2}{K-1}$ to compensate for the differing number of terms so that $J^{(1)}_{mean}$, $J^{(2)}_{var}$ and $J^{(2)}_{cov}$ have a similar influence on the result.

Implementation. In order to minimize (14) we apply a Quasi-Newton method. Setting $F := \nabla J$ such methods compute a local minimizer of J by iterating

$$\varepsilon^{(r+1)} = \varepsilon^{(r)} - A_r F(\varepsilon^{(r)}),$$

where A_r is an approximation of the inverse of the Hessian of J at $\varepsilon^{(r)}$ which has to fulfill the Quasi-Newton condition $A_r\big(F(\varepsilon^{(r+1)}) - F(\varepsilon^{(r)})\big) =$

$\varepsilon^{(r+1)} - \varepsilon^{(r)}$. We use the BFGS formula, see [13, 14, 15, 16], to produce the matrices A_r. The described Quasi-Newton method is implemented in the Medium-Scale algorithm of the `fminunc` function provided by Matlab (Version: R2008b, www.mathworks.org). The gradients of our higher order terms are given by

$$\begin{aligned}
\nabla J^{(1)}_{\text{mean}}(\varepsilon) &= \frac{1}{\sigma^2}\left((\mu_k - \mu)\,\mathbf{1}_{\widetilde{N}}\right)_{k=1}^{K}, \\
\nabla J^{(2)}_{\text{var}}(\varepsilon) &= \frac{1}{\sigma^4}\Big(\big(\frac{1}{\widetilde{N}}\|\tilde{\varepsilon}_k - \mu\|^2 - \sigma^2\big)\,(\tilde{\varepsilon}_k - \mu)\Big)_{k=1}^{K}, \\
\nabla J^{(2)}_{\text{cov}}(\varepsilon) &= \frac{1}{\widetilde{N}\sigma^4}\Big(\sum_{\substack{l=1\\ l\neq k}}^{K}\langle\tilde{\varepsilon}_k - \mu, \tilde{\varepsilon}_l - \mu\rangle\,(\tilde{\varepsilon}_l - \mu)\Big)_{k=1}^{K}.
\end{aligned}$$

The minimization procedure is initialized with $\varepsilon^{(0)} := g$ as a first guess of the noise vector ε. Thus, our reconstruction of the noisefree signal is first set to be $f = 0$.

Quantitative evaluation. To be able to quantify the quality of the results we use synthetic signals so that the original data vectors f are known and we can directly compare them to the obtained results $\hat{f}$. As a simple quality measure we apply the mean square error (MSE) criterion

$$\text{MSE}_{f,\hat{f}} = \frac{1}{N}\sum_{i=1}^{N}\left(f_i - \hat{f}_i\right)^2. \tag{15}$$

For our experiments we use the MAP-TM and MAP-TV approaches commented on in Section 1 as reference methods. To choose the involved regularization parameters λ in an optimal way, we perform a brute force search and use the parameter for which the reconstruction result produces the smallest MSE.

Experiments. By the subsequent examples we will show that our new approach is particularly well suited for signals with a lot of oscillations, since in these cases it is particularly hard to distinguish the signal parts from the noise and the usually applied smoothness terms fail to reconstruct the signals. Our first example in Figure 2 shows a signal constructed by adding three sine signals of different frequencies and amplitudes. To obtain its noisy version, this signal has been corrupted by additive Gaussian noise with $\mu = 0$

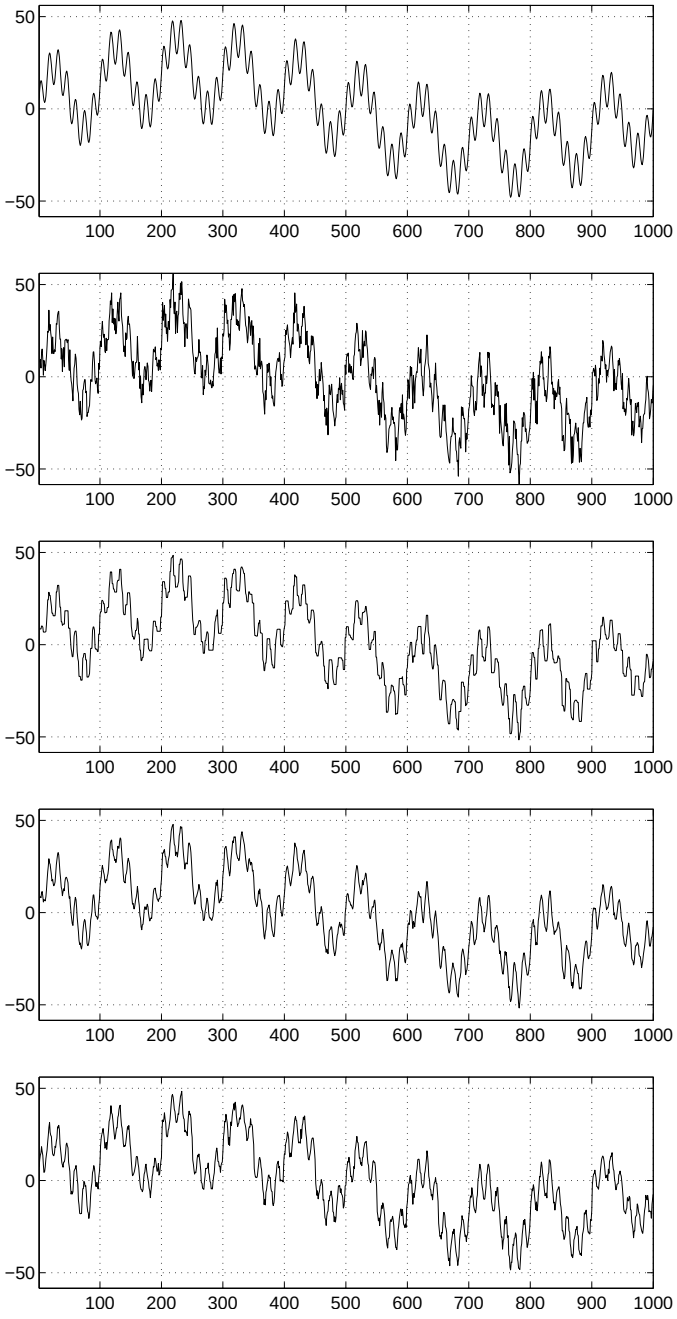

Figure 2: *From top to bottom:* Signal 1, noisy signal, reconstruction by MAP-TV ($\lambda = 3.8$), MAP-TM ($\lambda = 1.1$) and HOS ($\lambda = 0.0001$).

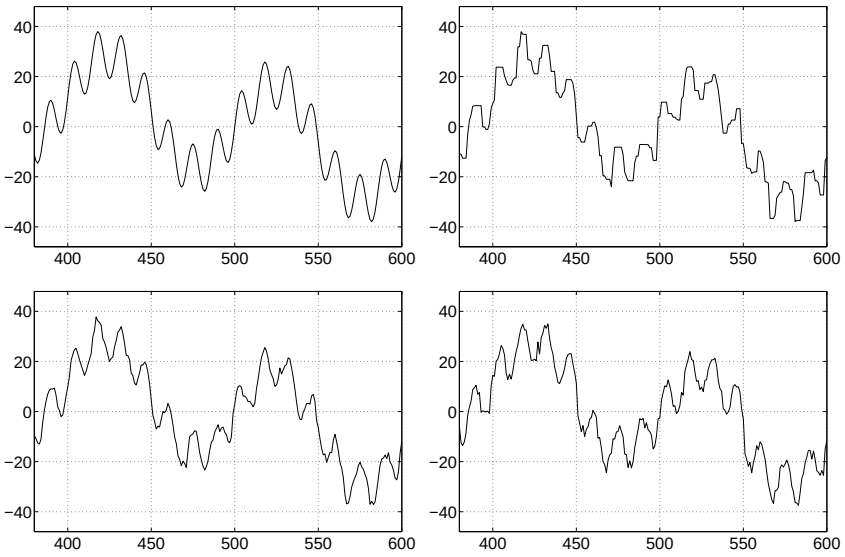

Figure 3: *From top left to bottom right:* Zoom into signal 1 as well as into MAP-TV, MAP-TM and HOS results.

and $\sigma^2 = 20$. As displayed in Figures 2 and 3 our HOS method yields a good reconstruction of the original signal. Table 1 shows that the MSE of this result is significantly better than the ones of the MAP-TM and MAP-TV results. Moreover, our HOS method produces a much better reconstruction $\hat{\varepsilon}$ of the original noise vector as illustrated in Figure 4. Here, the histogram of our noise values approximates well the shape of the density function of the noise, and the sample mean $\widetilde{\mu}$ and sample variance s^2 defined by

$$\widetilde{\mu} := \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \qquad \text{and} \qquad s^2 := \frac{1}{N}\sum_{i=1}^{N} (\varepsilon_i - \widetilde{\mu})^2$$

are very close to the original parameters μ and σ^2. Furthermore, if we compute the matrices $\left(\frac{1}{N}\langle \widetilde{\varepsilon}_k - \mu, \widetilde{\varepsilon}_l - \mu\rangle\right)_{l,k=1}^{K}$ from our reconstructed noise signal as done in Figure 8 (left), we see that it approximates well a diagonal matrix with diagonal entries $\sigma^2 = 20$. Thus, the subvectors of our reconstructed noise signal are uncorrelated and have the right variances.
These results are also confirmed by our second example displayed in Figures 5, 6, 7 and 8 (right). Here, a highly oscillating signal with several jumps is used which has been corrupted by additive Gaussian noise with $\mu = 0$ and $\sigma^2 = 80$.

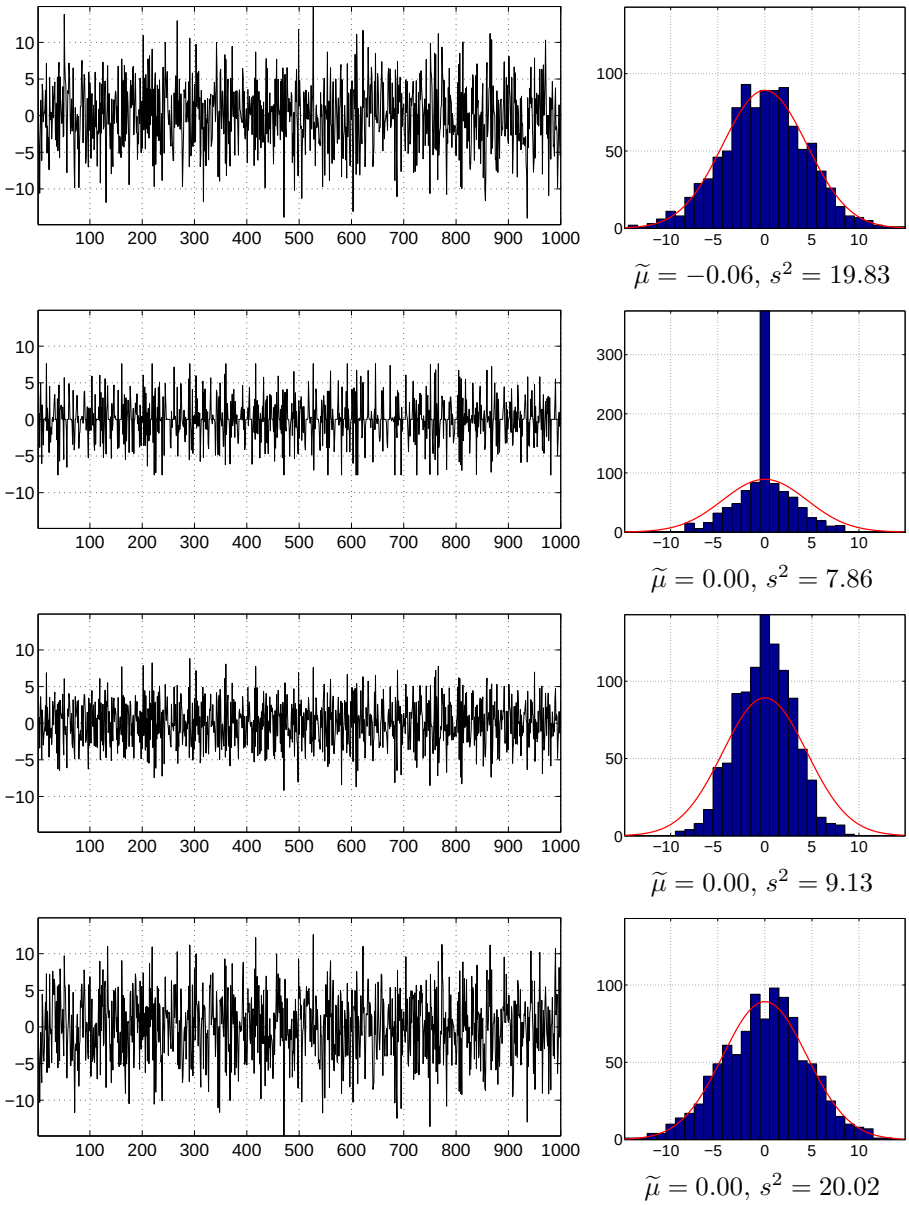

Figure 4: *Left, top to bottom:* True noise signal used for corrupting signal 1, reconstructed noise by MAP-TV, MAP-TM as well as HOS. *Right:* Histograms of the noise values with corresponding sample mean and sample variance. The red curve shows a by N scaled Gaussian with $\mu = 0$ and $\sigma^2 = 20$ for comparison.

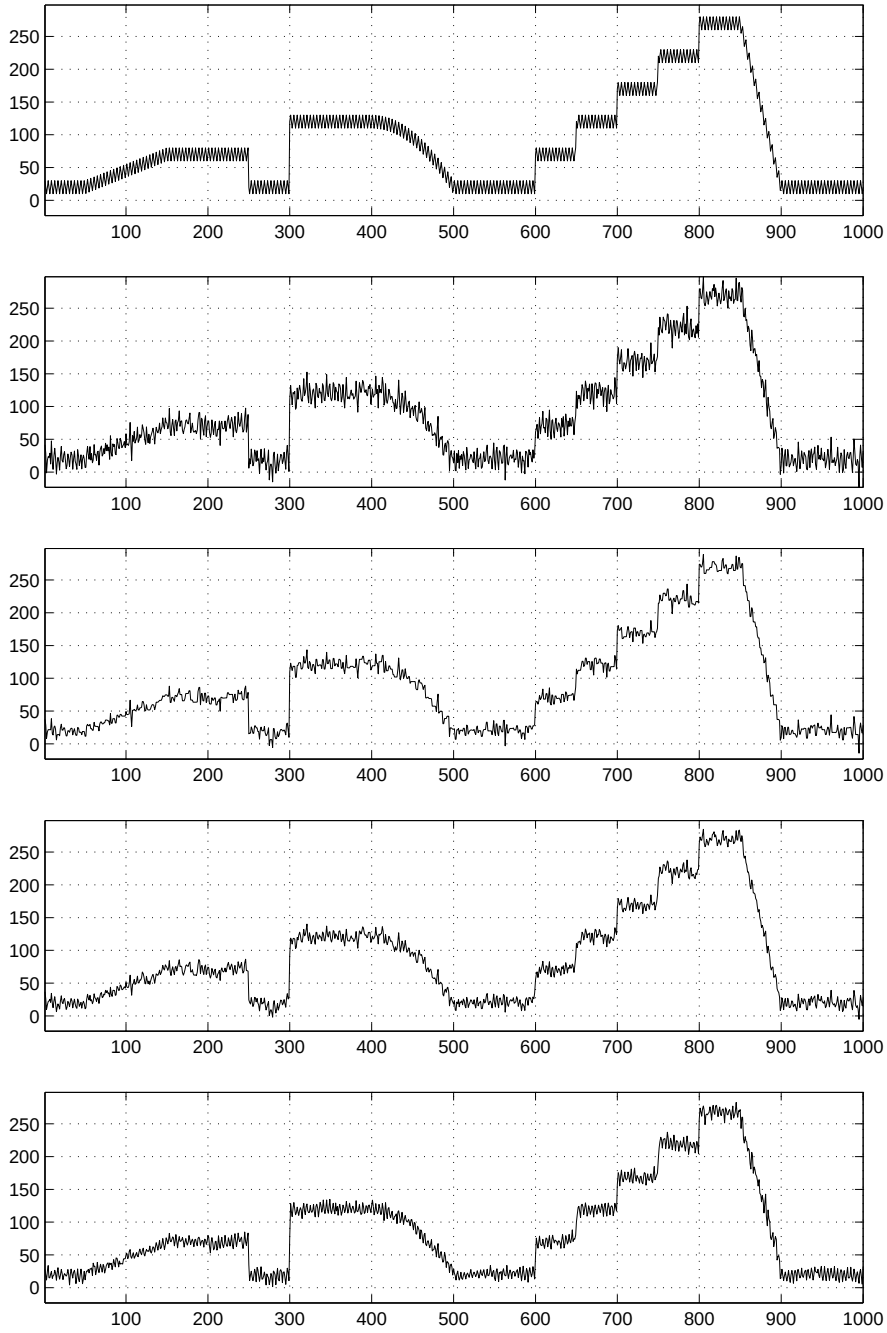

Figure 5: *From top to bottom:* Signal 2, noisy signal, reconstruction by MAP-TV ($\lambda = 4.6$), MAP-TM ($\lambda = 0.46$) and HOS ($\lambda = 0.00005$).

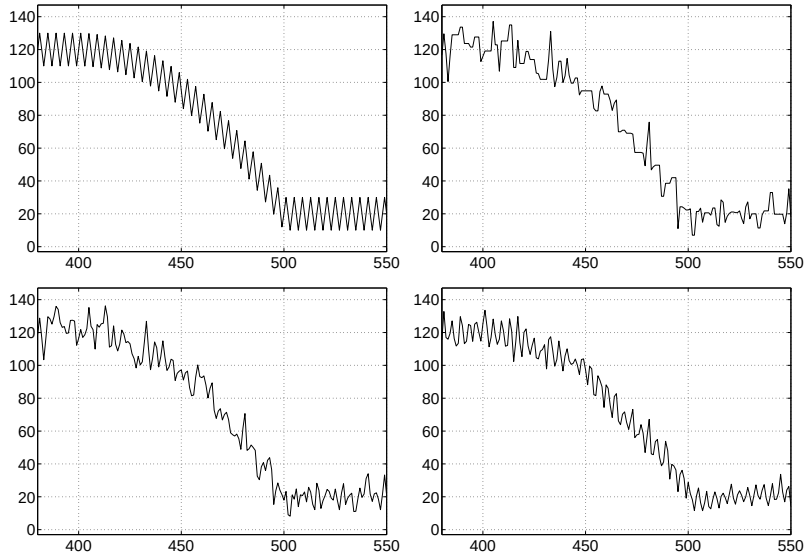

Figure 6: *From top left to bottom right:* Zoom into signal 2 as well as into MAP-TV, MAP-TM and HOS results.

Signals	signal 1	signal 2
Noisy signal	20.00	79.98
MAP-TV	9.52	52.64
MAP-TM	6.62	46.25
HOS	5.39	31.13

Table 1: Average MSE of the denoising results for 3000 different noisy realizations of the initial signals.

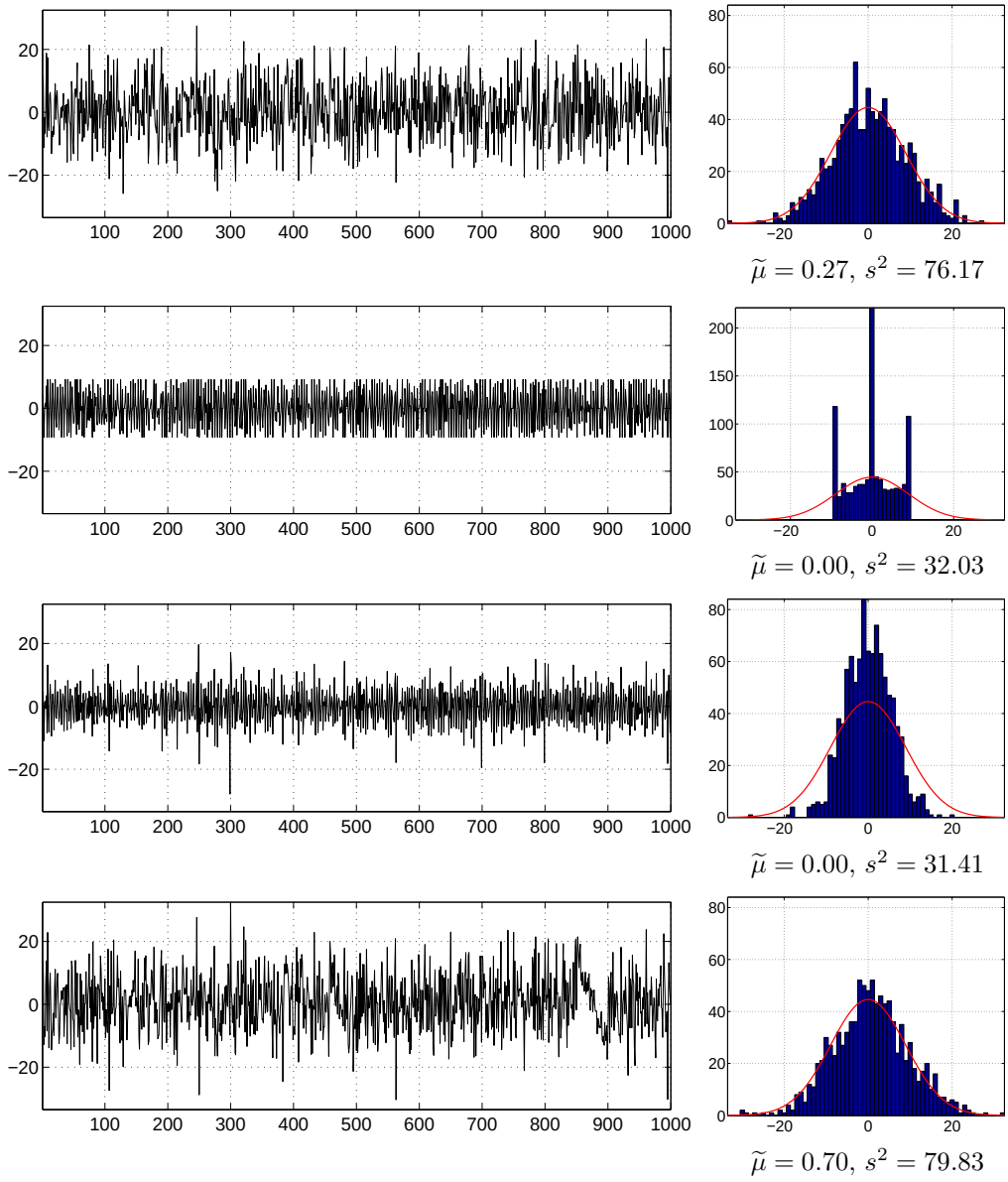

Figure 7: *Left, top to bottom:* True noise signal used for corrupting signal 2 and reconstructed noise signals by MAP-TV, MAP-TM as well as HOS. *Right:* Histograms of the noise values with corresponding sample mean and sample variance. The red curve shows a by N scaled Gaussian with $\mu = 0$ and $\sigma^2 = 80$ for comparison.

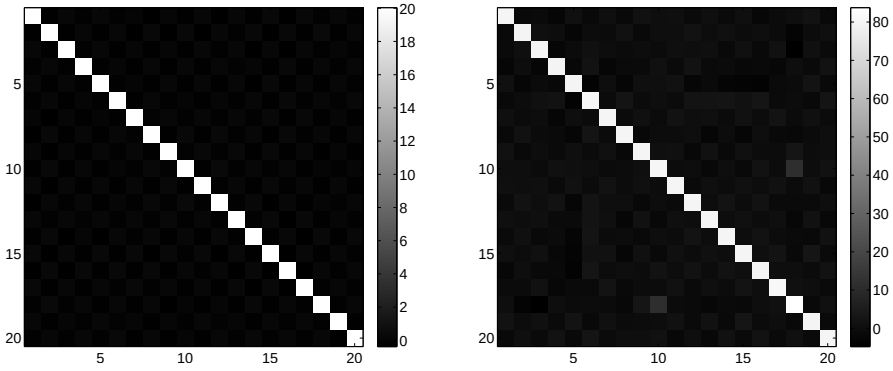

Figure 8: *Left to right:* Estimated covariance matrices $\left(\frac{1}{N}\langle\widetilde{\varepsilon}_k - \mu, \widetilde{\varepsilon}_l - \mu\rangle\right)_{l,k=1}^{K}$ for the reconstructed noise signals by our HOS approach displayed in Figures 4 and 7, resp. As expected, we obtain nearly diagonal matrices with diagonal entries around σ^2. The means of the absolute values of the off-diagonal entries are 0.23 (left) and 1.20 (right).

4. Conclusions

We have shown that the standard maximum likelihood estimation approach for denoising signals can be generalized by introducing an additional transformation of the random variables modeling the noise. This transformation allows to consider also pixel correlations within the noise vectors and helps to obtain a reconstructed noise vector, which resembles the statistical properties of the assumed noise model. The transformation of our choice leads to a nonconvex minimization problem. A local minimizer of the functional was computed by a BFGS Quasi-Newton method. In order to evaluate the capability of our new approach, we performed feasibility tests on different signals. These experiments showed that especially in cases where the signal consists of high frequency components, our approach allows to recover even fine structures in the data. These results give hope that our HOS approach allows for more detailed and accurate reconstructions compared to standard techniques. As a topic of future research, different, more complex transformations including for example multiple splittings could be considered. Moreover, we aim to integrate suitable, more sophisticated regularization terms modeling the a priori knowledge about the data vector f in our reconstruction procedure. To this purpose, suitable minimization methods have to be found to be able to handle the resulting possibly nondifferentiable, nonconvex minimization problems.

[1] A. Tikhonov, Regularization of incorrectly posed problems, Soviet Mathematics Doklady 4 (1963) 1624–1627.

[2] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1992) 259–268.

[3] D. Boes, F. Graybill, A. Mood, Probability and statistical inference, 3rd Edition, McGraw-Hill, New York, 1974.

[4] W. Press, B. Flannery, S. Teukolsky, W. Vetterling, Numerical Recipes in FORTRAN: The Art of Scientific Computing, Cambridge University Press, Cambridge, England, 1992.

[5] H. W. Engl, A. Hofinger, S. Kindermann, Convergence rates in prokhorov metric for assessing uncertainty in ill-posed problems, Journal of Inverse Problems 21 (1) (2005) 399–412.

[6] A. Hofinger, H. Pikkarainen, Convergence rates for linear inverse problems in the presence of additive normal noise, Journal of Stochastic Analysis and Applications 27 (2) (2009) 240–257.

[7] R. G. Clapp, Multiple realizations and data variance: Successes and failures, Tech. Rep. 113, Stanford Exploration Project (2003).

[8] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, 2005, pp. 60–65.

[9] K. Frick, P. Marnitz, A. Munk, Statistical multiresolution estimation in imaging: Fundamental concepts and algorithmic framework, eprint arXiv:1101.4373 (2011).

[10] J. Jacod, P. Protter, Probability Essentials, 2nd Edition, Springer, 2004.

[11] N. Mukhopadhyay, Probability and statistical inference, Marcel Dekker, New York, 2000.

[12] R. J. Serfling, Approximation Theorems of Mathematical Statistics, John Wiley & Sons, Inc., 1980.

[13] C. G. Broyden, The convergence of a class of double-rank minimization algorithms, IMA Journal of Applied Mathematics 60 (1) (1970) 76–90.

[14] R. Fletcher, A new approach to variable metric algorithms, The Computer Journal 13 (3) (1970) 317–322.

[15] D. Goldfarb, A family of variable-metric updates derived by variational means, Mathematics of Computing 24 (109) (1970) 23–26.

[16] D. F. Shanno, Conditioning of quasi-newton methods for function minimization, Mathematics of Computing 24 (111) (1970) 647–656.