RECONSTRUCTION OF HIGHLY STRUCTURED IMAGE BY ENTROPY OPTIMIZATION

Zhaohui H. Sun

Kitware Inc., Clifton Park, NY, USA 12065

ABSTRACT

We propose an image reconstruction approach for highly structured image based on entropy optimization, extending the idea of sparsity with minimal number of non-zero elements to the concept of minimum entropy and simple structure. Minimum entropy signals have unique characteristics of low uncertainty, simple structure, and high concentration, and are usually sufficiently sparse. However, not all low-entropy signals can be linearly sparsified. Compared to a sparse signal in L_0 or L_1 norm having one dominant mode centered around 0, a highly structured signal with minimum entropy tends to be concentrated in one or more dominant modes, not necessarily around 0. We study histogram entropy for finite discrete signal and vector entropy for continuous real or complex signal, and propose two algorithms to seek the optimal signal with minimum entropy or low theoretic dimension, subject to the reconstruction constraint. Experiments are carried out on 1-D finite discrete signal, 2-D halftone, industrial and medical images to demonstrate the efficacy.

1. INTRODUCTION

It has been shown that an inherently sparse signal can be recovered from few measurements with high probability through compressive sensing [1, 2, 3], i.e., nonuniform sampling followed by sparse reconstruction through L_0 or L_1 optimization. Consider a linear system $\mathbf{y} = \mathbf{A}\mathbf{x}$, with the signal $\mathbf{x} = (x_1, \cdots, x_N)^T \in \Re^{N\times 1}$, the transform matrix $\mathbf{A} \in \Re^{M\times N}$, and the measurement $\mathbf{y} = (y_1, \cdots, y_M)^T \in \Re^{M\times 1}$. When the linear system is under-constrained ($M < N$), there are an infinite number of signals $\mathbf{x}$ which could yield the same measurement $\mathbf{y}$ [4]. However, when a sparse signal $\mathbf{x}$ has many zero elements, with no numerical contribution to the measurements, it is possible to recover the signal from significantly fewer measurements ($M << N$).

Most visual signals (e.g. image and video) have specific structures, such as piece-wise linear and energy concentration, and can be sparsified in the space domain and transform (Fourier and wavelet) domains. Since only a small fraction of the coefficients are non-zero, it is possible to significantly reduce the number of measurements. This formulation dramatically changes the work flow of signal capture, compression, representation, and reconstruction, leading to new imaging devices and methods, such as the single pixel camera [5], fast MRI scanner [6], and CT machines for medical, industrial and security applications. In addition, sparse reconstruction also plays an interesting role in feature selection, especially in machine learning from a small number of samples.

The key ingredients are random sampling and sparse reconstruction algorithms. The reconstruction problem is formulated as L_0 optimization [1, 2] for $p = 0$,

$$\min \|\mathbf{x}\|_p \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon, \tag{1}$$

where the L_0 norm counts the number of non-zero elements. Optimization (1) finds the sparest vector $\mathbf{x}$ with as few non-zero elements as possible. The reconstruction constraint $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon$ indicates the sampled estimate $\mathbf{A}\mathbf{x}$ has minimal Euclidean distance to the measurement $\mathbf{y}$, i.e. the difference is within a L_2 ball with a small radius ϵ. When $\epsilon = 0$, it is the exact reconstruction constraint. Since L_0 optimization (1) is NP hard and has combinatorial complexity, it has been relaxed to L_1 optimization for $p = 1$, where $\|x\|_1 = \sum_{n=1}^{N} |x_n|$. It has been shown in [7, 1, 2] and others that the L_1 optimization has high probability of success on sufficiently sparse vector $\mathbf{x}$. Problem (1) has been solved by matching pursuit [8], basis pursuit [9], the interior point algorithms for linear program (LP) and second-order cone program (SOCP) [10], and the total variation algorithm [1]. In practice, it is the reconstruction details that make a difference, such as the computation speed, accuracy, and optimization convergence.

In this paper, we propose entropy optimization (2,3) as an alternative to the L_0 and L_1 optimization in (1) and show how to find the highly structured solution through an iterative algorithm using nonlinear optimization by closed-form gradient descent. We study constrained histogram entropy minimization for signals drawing values from a finite discrete set, e.g. finite integer or quantized real set, which is a much stronger prior. We also study vector entropy minimization for continuous real or complex signals, through conjugate gradient methods, where a scaling function is introduced to adapt to the estimate over iterations.

Entropy is a measure of the inherent uncertainty, concentration, localization, and structure. By entropy minimization, we seek a highly structured vector $\mathbf{x}$ that satisfies the reconstruction constraint. Note this is somewhat equivalent but different from seeking a vector with few non-zero elements. Although sparsity implies simple structure with low entropy (a

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

small fraction of non-zero elements), not all low-entropy signals can be linearly sparsified. When $\mathbf{x}$ has a single dominant mode centered around 0, the entropy optimization and L_0/L_1 optimization all converge to the same unique solution. The difference arises when the signal has multiple dominant modes, such as the black/white states in the halftone images in Fig. 2. Since L_0/L_1 optimization is not effective for simple signals with multiple modes, we argue that minimum entropy is a more fundamental signal prior. This makes an interesting connection to coding theory and has the potential of extending compressive sensing to more generic signals (not necessary linearly sparse), further decreasing the number of measurements (M), and increasing the Equivalence Breakdown Point (EBP) [7]. Entropy minimization has been used before, e.g. for base selection [11] and data segmentation [12]. To our best knowledge, it is the first time that sparse reconstruction is cast as an entropy minimization framework, and "sparsity" is interpreted in the context of simple structure and high concentration for highly structured signal recovery.

2. PROBLEM FORMULATION

The problem is to find *highly structured* signal $\mathbf{x} \in \Re^N$, which yields the observed measurement $\mathbf{y} \in \Re^M$ under linear transform $\mathbf{A} \in \Re^{M\times N}$. When signal structure is measured by entropy [13], it is written as entropy minimization

$$\min H(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 \le \epsilon, \qquad (2)$$

where $H(\mathbf{x})$ is the entropy of signal $\mathbf{x}$.

The entropy minimization in (2) is quite different from the L_0 and L_1 optimization in (1), which seeks sparse signal with one dominant mode centered around 0. Simple structure is a more general concept than having minimal number of non-zero elements. Entropy minimization works for signals with multiple modes (e.g. the black and white states in halftone image), therefore it is applicable to more general sparse signals. Although sparsity implies minimum entropy, not all signals with simple structure can be linearly sparsified. On the other hand, L_0/L_1 are metrics and L_1 is convex. The convex optimization is well studied, and many algorithms are readily available, such as LP and SOCP. In comparison, the entropy measure is not a metric. Entropy minimization requires nonlinear optimization and potentially has higher complexity. The global optimum is not always guaranteed.

The entropy minimization formulation (2) can be equivalently written as dimension reduction

$$\min D(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 \le \epsilon, \qquad (3)$$

where the theoretic dimension $D(\mathbf{x}) = exp\{H(\mathbf{x})\}$ is an exponential function of entropy [11], indicating the inherent dimensionality of the signal. It is no more than the number of unique symbols $0 \le D(\mathbf{x}) \le Q$. For sparse signals, it is much smaller than the signal length $Q << N$.

We use Lagrange multiplier to solve the constrained optimization problems (2)

$$\mathcal{E}(\mathbf{x}) \quad = \quad H(\mathbf{x}) + \lambda\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2, \qquad (4)$$

where $\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathcal{E}(\mathbf{x})$ is the optimal solution, and λ is the parameter to balance the entropy/dimension term and the reconstruction term. In the following, we show how to solve (4) by histogram entropy optimization for finite discrete signal in Section 3 and by vector entropy optimization for continuous real signal in Section 4.

3. HISTOGRAM ENTROPY OPTIMIZATION

Consider a signal $\mathbf{x}$ that takes values from an ordered finite discrete set $\mathbf{Q} = [s_1, \cdots, s_Q]$, where $Q << N$, $s_i \le s_j$ for $i < j$. The examples include $[0, 1]$, $[\pi, e]$, and $[0, \cdots, 255]$. A sparse signal is a special case with $0 \in \mathbf{Q}$. This is a much stronger signal prior and the situation arises frequently in practice. For example, the pixels of a halftone image take only binary values of $[0, 1]$, and an 8-bit image takes a total of 256 potential values $[0, \cdots, 255]$.

The entropy in (2) is chosen as Shannon entropy, $H(\mathbf{x}) = -\sum_{i=1}^{Q} p(s_i)\log p(s_i)$. It is a nonlinear function of signal $\mathbf{x}$, where $p(s_i) \in [0, 1]$ is the empirical probability with $p(s_i) \ge 0$ and $\sum_i p(s_i) = 1$. The histogram $\mathbf{h}(\mathbf{x}) = (h_1, \cdots, h_Q)$ is defined as $h_i(\mathbf{x}) = \sum_{n=1}^{N} \delta(x_n - s_i)$, $i = 1, \cdots, Q$, where Dirac delta $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. The histogram counts the number of elements in $\mathbf{x}$ taking value of s_i. The probability p is then defined as the chance of an element taking a particular value s_i, $p_i(\mathbf{x}) = \frac{h_i(\mathbf{x})}{\sum_i h_i(\mathbf{x})}$.

To minimize entropy $H(\mathbf{x})$, one needs to make the probability of certain events either very small (close to 0) or very large (close to 1). As the elements of $\mathbf{x}$ are concentrated at some isolated symbols, the signal dimensionality $D(\mathbf{x})$ is greatly reduced. To optimize (2,3), x_n is swapped from one value to another in the finite set $\mathbf{Q}$ while meeting the reconstruction constraint. The exhaustive search has a complexity of $O(Q^N)$, and efficient algorithms can be derived from matching pursuit or basis pursuit.

We use gradient decent [14] to solve the constrained entropy minimization problem (4). The idea is to test the value of element x_n in the ordered discrete finite set, away from the gradient direction in the solution space, until the minimum is reached. The gradient of the cost function (4) is

$$\frac{\partial\mathcal{E}(\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial H(\mathbf{x})}{\partial\mathbf{x}} + 2\lambda\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{y}) \qquad (5)$$

To facilitate the gradient computation, we replace the Dirac delta with a differentiable kernel $\kappa_\sigma(x) = \exp\{-\frac{x^2}{2\sigma^2}\}$, and define the soft histogram as

$$h_i(\mathbf{x}) = \sum_{n=1}^{N} \kappa_\sigma(x_n - s_i) = \sum_{n=1}^{N} e^{-\frac{(x_n-s_i)^2}{2\sigma^2}} \qquad (6)$$

The kernel takes the maximum value of 1 when $x_n = s_i$ and a very small value (close to 0) otherwise. The kernel standard deviation σ is chosen as a fraction of the minimal adjacent symbol difference, e.g., $\frac{1}{3}\min_i |s_{i+1} - s_i|$.

The histogram entropy gradient in the closed-form format can then be derived as

$$\begin{aligned}\frac{\partial H(\mathbf{x})}{\partial x_n} &= \frac{\partial}{\partial x_n}(-\sum_{i=1}^{Q} p_i(\mathbf{x}) \log p_i(\mathbf{x})) \\ &= \frac{1}{Z\sigma^2}\sum_{i=1}^{Q}((1+\log p_i(\mathbf{x}))(x_n - s_i)e^{-\frac{(x_n-s_i)^2}{2\sigma^2}}),\end{aligned} \quad (7)$$

where $p_i(\mathbf{x})$ is the histogram probability, $Z = \sum_i h_i(\mathbf{x})$ is the normalization factor, n is the signal index $\mathbf{x} = (x_1, \cdots, x_N)$, and i is the finite discrete set index $\mathbf{Q} = [s_1, \cdots, s_Q]$.

The nonlinear optimization problem is solved by an iterative gradient descent method in the finite discrete solution space. At a specific iteration, the cost function $\mathcal{E}(\mathbf{x})$ in (4) and the cost gradient $\mathcal{E}'(\mathbf{x})$ in (5) are evaluated. The element x_n having the largest absolute gradient is tested with the symbols away from the gradient direction. It is a 1-D search in the finite discrete set for symbols either larger or smaller than x_n, depending on the gradient direction. The search continues until a symbol with a lower cost is found, $\mathcal{E}(s_i) < \mathcal{E}(x_n)$, then update $x_n = s_i$ and continue the iteration. If all symbols have been exhausted in the single-sided 1-D search, move on to the signal element with the second largest absolute gradient. The iteration continues and it should stop in finite steps. To avoid local minimum, the procedure is repeated multiple times from different random starting points, until the reconstruction constraint is satisfied. If the signal structure is simple enough, it usually converges in a few iterations.

4. VECTOR ENTROPY OPTIMIZATION

When signal $\mathbf{x} \in \Re^{N\times 1}$ takes continuous real or complex values, (4) is solved by vector entropy optimization. The empirical probability is defined as

$$p(x_n) = \frac{1}{Z(\mathbf{x})}\rho_\alpha(x_n), \quad (8)$$

where $Z(\mathbf{x}) = \sum_{n=1}^{N} \rho_\alpha(x_n)$ is the normalization parameter, and $\rho_\alpha(x_n)$ is a scaling function with bandwidth parameter α and $\rho_\alpha(0) = 0$. The scaling function $\rho_\alpha(x_n)$ maps x_n to probability $p(x_n)$. Some basic requirements are zero origin $\rho_\alpha(0) = 0$, symmetric $\rho_\alpha(-x) = \rho_\alpha(x)$, and concave shape $\rho''_\alpha(x) \le 0$ for $x > 0$. In other words, large changes happen around 0, and the growth rate drops as x increases. There are many functions in this family. For convenience, we choose a specific form $\rho_\alpha(x) = 1 - e^{-\frac{|x|^2}{2\alpha^2}}$, with gradient $\rho'_\alpha(x) = \frac{x}{\alpha^2}e^{-\frac{|x|^2}{2\alpha^2}}$. The bandwidth parameter α controls the function shape. The zero elements are mapped to zero probability with no contribution to the vector entropy, since $p(x_n)\log p(x_n) = 0$ when $x_n = 0$. Therefore, vector entropy minimization works for the sparse signals with few non-zero elements.

The conjugate gradient method and 1-D line search (similar to [6]) are used to solve the constrained vector entropy minimization problem in (4). The key is to find the closed-form gradient, which can be derived as

$$\begin{aligned}\frac{\partial H(\mathbf{x})}{\partial x_n} &= -(1+\log p(x_n))\frac{\partial p(x_n)}{\partial x_n} \\ &= -\frac{1}{Z(\mathbf{x})}(1+\log p(x_n))(1-p(x_n))\rho'_\alpha(x_n).\end{aligned} \quad (9)$$

5. EXPERIMENTAL RESULTS

We carry out experiments on 1-D signal and 2-D halftone, medical and industrial images to validate the constrained entropy minimization algorithms, and compare the results by *l1magic* [1, 3] and *csMri* [6, 3] using L_1 optimization.

First consider the 1-D signal in Fig. 1(a), with values taken from a finite discrete set $x(n) \in \mathbf{Q} = [-3.5, 0, 9]$. It has relatively simple structure and a low entropy of 0.962, with 40 random spikes in a step transition from 9 to -3.5. The signal is also perturbed by zero mean Gaussian noise with deviation of 0.5, as $\tilde{x}(n)$ in (b). The signal itself and its Fourier coefficients in (c) are not sufficiently sparse in L_0. Random matrices $\mathbf{A} \in \Re^{M\times N}$ with full rank are generated and the projections $\mathbf{y} = \mathbf{A}\mathbf{x}$ are taken as the measurements. The histogram entropy minimization algorithm perfectly recovers the original signal $x^*_{entropy}(n) = x(n)$ from 120 (60%) random measurements of $x(n)$ in (a). Repeating the same algorithm on 120 random measurements of the noisy signal $\tilde{x}(n)$ in (b) also yields the perfect reconstruction $\tilde{x}^*_{entropy}(n) = x(n)$. The strong prior of finite discrete constraint greatly reduces the possible combinations and the solution space. In comparison, L_1 optimization has difficulty in signal reconstruction in (d,e). Using *l1eq* in *l1magic*, the reconstruction of $x_{l1eq120}(n)$ from 120 measurements of $x(n)$ in (d) is far from the true signal. From noisy measurements, algorithm *l1qc* in *l1magic* yields reconstruction (e) even with 180 (or 90%) measurements $\tilde{x}'_{l1eq180}(n)$. The study is repeated for 50 times and the trend is clear. Histogram entropy minimization can recover highly structured signals without explicit sparsification and require fewer measurements, even when the signals are not sparse in L_0.

Next we apply entropy minimization to 2-D halftone image reconstruction. A halftone image uses reduced tone scales to render a continuous tone image with nice visual effect. The image has simple structure, as the pixel has binary modes of being black or white. However, it is usually not sufficiently sparse, and is also difficult to be sparsified due to the high frequency components. Such a highly structured image poses great challenges to L_1 optimization. Fig. 2(a) shows the

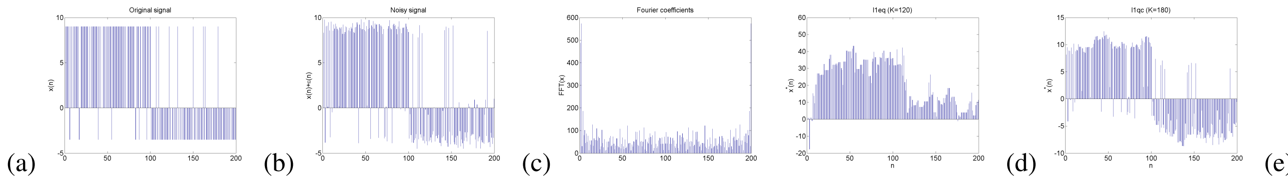

Fig. 1. Reconstruction of 1-D signal. Histogram entropy minimization can perfectly recover the highly structured signal $x(n)$ in (a) from 120 random measurements of the noise perturbed signal $\tilde{x}(n)$ in (b). The Fourier coefficients in (c) are not sparse. In comparison, L_1 optimization has difficulty to recover $x_{l1eq120}(n)$ using 120 measurements in (d) and $\tilde{x}'_{l1qc180}(n)$ using 180 measurements in (e) by *l1qc* algorithm of *l1magic*.

Fig. 2. (a) The original halftone image with 48% bright and 52% dark pixels. (b) The Fourier coefficients. (c) L_1 optimization from 2500 measurements. (d,e) Histogram entropy optimization using 2374 and 2500 (61%) measurements.

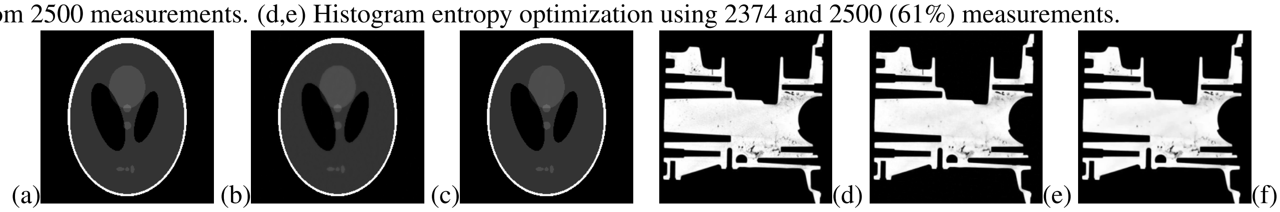

Fig. 3. Reconstruction of the Shepp-Logan phantom (a,b,c) and the CT image of an auto part (d,e,f). (a,d) The original images. (b,e) Recovery by vector entropy minimization. (c,f) Recovery by total variation L_1 minimization (csMri).

halftone Lena image with a dimensionality of 64×64. It has 1985 (48%) bright pixels (ones) and 2111 (52%) dark pixels (zeros), not sparse enough. The 2-D Fourier coefficients in (b) are not sparse either. The L_1 optimization by *l1eq* with 2500 random measurements yields (c). Even when 85% measurements, it still cannot recover the image properly. In comparison, the histogram entropy minimization can partially recover the image in (d) from 2374 random measurements. When the number of random measurements is increased to 2500 (61%), the original halftone image can be perfectly recovered in (e). For signals with two or more dominant modes, such as the binary modes in halftone image, constrained entropy minimization is a better choice than L_1 optimization.

At last, we test out the constrained vector entropy minimization algorithm in Section 4 on the medical and industrial images, and compare the results with those using L_1 optimization in *csMri*. When the image or the transform domain (Fourier or Wavelet) coefficients are sufficiently sparse, the vector entropy minimization provides an alternative path to the optimal solution. Fig. 3 shows the results on the the Shepp-Logan phantom and the CT image of an auto part [15]. The original images, shown in (a,d), are piece-wise smooth, and the energy is concentrated at the low frequency coefficients. A total of 35% random measurements are made in the Fourier domain, with high sampling density around the low frequencies [6]. (b,e) show the reconstruction results by the vector entropy minimization after 20 iterations, and (c,f) show the results by the total variation L_1 minimization in *csMri* [6]. The results are comparable. When the signal is sufficiently sparse, they all converge to the same solution from the random measurements.

6. CONCLUSION

We have proposed entropy minimization as a strong prior for reconstruction of highly structured image, and extended sparsity, in terms of limited number of non-zero coefficients, to the concept of minimal entropy, simple structure, and low theoretic dimension. Entropy minimization serves as an alternative to the L_0/L_1 optimization which are widely used in compressive sensing, so the same great idea can be applied to more generic signals. Two reconstruction algorithms have been proposed, including the histogram and vector entropy minimization for discrete finite signal and continuous real or complex signal, respectively. The proposed method is particularly suitable for highly structured signals and images with multiple dominant modes.

Acknowledgement Part of the work was done at GE Global Research Center, Niskayuna, New York.

7. REFERENCES

[1] Emmanuel Candes, Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[2] David Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[3] Compressive Sensing Resources, "http://www.dsp.ece.rice.edu/cs," .

[4] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins University Press, Baltimore, Maryland, 2013.

[5] Marco Duarte, Mark Davenport, Dharmpal Takhar, Jason Laska, Ting Sun, Kevin Kelly, and Richard Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, March 2008.

[6] Michael Lustig, David Donoho, and John M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, December 2007.

[7] David Donoho, "For most large underdetermined systems of linear equations, the minimal L-1 norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, June 2006.

[8] Stephane Mallat and Zhifeng Zhang, "Matching pursuit with time frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[9] Scott Shaobing Chen, David Donoho, and Michael Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[10] Emmanuel Candes and Terence Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203 – 4215, December 2005.

[11] Ronald R. Coifman and Mladen Victor Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, March 1992.

[12] David L. Donoho, "On minimum entropy segmentation," Tech. Rep., Stanford University, 1994.

[13] T. Cover and J. Thomas, *Elements of information theory*, John Wiley and Sons, New York, NY, 1991.

[14] W. H. Press, S. A. Teulolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, Cambridge University Press, 2007.

[15] D. Dastarac, "Industrial computed tomography: control and digitalisation," in *International Symposium on Computerized Tomography for Industrial Applications and Image Processing in Radiology*, Berlin, Germany, 1999.