SPARSE CODE SHRINKAGE BASED ON THE NORMAL INVERSE GAUSSIAN DENSITY MODEL

Robert Jenssen, Tor Arne Øigård, Torbjørn Eltoft and Alfred Hanssen

Department of Physics
University of Tromsø
N - 9037 Tromsø, Norway

ABSTRACT

In this paper we introduce the recent normal inverse Gaussian (NIG) probability density as a new model for sparsely coded data. The NIG density is a flexible, four-parameter density, which is highly suitable for modeling unimodal super-Gaussian data.

We demonstrate that the NIG density provides a very good fit to the sparsely coded data, obtained here via an independent component analysis (ICA) transform of the observations. In image denoising, we utilize this new density by developing a NIG-based maximum a posteriori estimator of a sparsely coded image corrupted by white Gaussian noise. The estimator acts as a shrinkage operator on the noisy components in the sparse domain. We demonstrate the technique by an image denoising experiment.

1. INTRODUCTION

Wavelet transformed images, or images in an ICA representation, typically have super-Gaussian probability density functions (pdf's), i.e. they have positive normalized kurtosis. We refer to such representations as sparse codes, since the components of the representation only rarely deviate significantly from zero.

In the recent years sparse coding has been exploited in image denoising, by a "coring" [1], or a "shrinkage" [2, 3] technique. Small amplitude values are thought to originate from zero-valued components influenced by noise, and are suppressed, while large values are preserved.

Common for these techniques are the need for a parameterized pdf model for the super-Gaussian components in the transform domain. A classical sparse density is the Laplacian density [4]. This is a one-parameter density, thus it is unable to model different degrees of kurtosis for a given variance. The two-parameter generalized Laplacian density [1] represents a modification of the Laplace density. This is a zero mean, symmetric density whose parameters are directly related to the second and fourth order moments. Other models were proposed by Hyvärinen in [3], and the author referred to these models as mildly sparse and very sparse densities. These are two-parameter, zero mean, symmetric models. The parameters are related to the second order moment, the expected absolute value, and in addition to the peak value of the density.

A proper statistical model should be flexible enough to provide a good fit to the data by having the ability to model various degrees of super-Gaussianity, and to take into account a possible skewness. In addition it should be possible to estimate the model parameters readily from the noisy observation.

In this paper we propose to use the recent normal inverse Gaussian (NIG) density [5] to model the super-Gaussian components. The NIG density has the flexibility that makes it capable of satisfying the requirements listed above, and in addition very fast cumulant based estimators for the four parameters of the density exist [6]. In the symmetric case, it can model data ranging from zero normalized kurtosis, i.e. the Gaussian distribution, to any positive valued kurtosis.

In this paper we develop a maximum a posteriori (MAP) denoising technique based on NIG modeled ICA-decomposed image data contaminated by white Gaussian noise. We show that the NIG density model fits the sparsely coded data well, and we provide an example of the method applied to a real noisy image.

2. MAXIMUM A POSTERIORI ESTIMATION OF SPARSE CODED SIGNAL

2.1. Sparse coding by ICA

In linear sparse coding we search for an $(n \times m)$ matrix $\mathbf{W}$ which transforms a set of observations $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ into a new representation $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$, making the components of $\mathbf{s}$ as sparse as possible. Hence,

$$\mathbf{s} = \mathbf{W}\mathbf{x}, \tag{1}$$

and $\mathbf{x}$ can be considered a m-dimensional random vector.

The ICA transform, on the other hand, aims at making the components of $\mathbf{s}$ as jointly statistically independent as

possible [7]. This reduces to a search for uncorrelated components being as non-Gaussian as possible [3]. Thus, if we constrain the components s_i of the sparse code model (1) to be uncorrelated, and if the ICA transformed data are in fact super-Gaussian, ICA will obey the sparse code model. Since ICA transformed image data are super-Gaussian, sparse coding can be accomplished. In ICA, m and n are normally equal. This is referred to as the complete case.

When observing noisy data, sparse coding is used to separate the statistics of the noise free signal from the noise in the transform domain. In the case of a signal $\mathbf{x}$ corrupted by white Gaussian noise i.e. $\mathbf{n} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, where σ^2 is the noise variance, the result of the transform

$$\mathbf{y} = \mathbf{W}(\mathbf{x} + \mathbf{n}) = \mathbf{s} + \mathbf{v}, \tag{2}$$

will lead to a maximally sparse $\mathbf{s}$ and yet Gaussian $\mathbf{v}$. We assume that $\mathbf{n}$ is statistically independent of $\mathbf{x}$.

Let $\hat{\mathbf{s}}$ denote the estimate of $\mathbf{s}$, based on $\mathbf{y}$. By doing the inverse transformation of $\hat{\mathbf{s}}$ we obtain $\hat{\mathbf{x}}$ as the denoised version of $\mathbf{x}$.

In ICA, the transform matrix $\mathbf{W}$ is in general not orthogonal. Thus, the components of $\mathbf{v}$ will be correlated. If we require $\mathbf{W}$ to be orthogonal, then $\mathbf{v} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The orthogonalization of $\mathbf{W}$ can be accomplished by $\mathbf{W} \leftarrow \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1/2}$ [3].

To estimate $\mathbf{W}$ in (2) directly from noisy data has proven to be an inherently difficult task. Therefore, when taking noisy data to the ICA domain, we assume that $\mathbf{W}$ has been estimated from noise free data in advance. This is the drawback of using ICA to obtain a sparse code.

The advantage is that the components of the representation are thought to be mutually independent. Thus, the MAP estimation of $\mathbf{s}$ given $\mathbf{y}$, which generally requires knowledge of the joint density of the sparse components, $f(\mathbf{s})$, can be approximated by a componentwise factorized operation for each individual component.

2.2. Maximum a posteriori estimation

Consider now a single noisy component

$$y = s + n, \tag{3}$$

where $n \sim N(0, \sigma^2)$. Our task is now to estimate s given y by $\hat{s} = g(y)$. The MAP estimator $\hat{s}$ of s is the value of s which maximizes $f(s|y)$, the conditional density of s given y. This is called the a posteriori probability density of s. It can be expressed via Bayes formula as

$$f(s|y) = \frac{f(y|s)f(s)}{f(y)}, \tag{4}$$

where $f(y|s)$ is the conditional density of the observation y given s, $f(s)$ is the density of the sparse component and $f(y)$ is the density of the noisy observation. Maximizing the a posteriori density (4) is equivalent to maximizing the product $f(y|s)f(s)$ since $f(y)$ is independent of the parameter s.

Furthermore,

$$f(y|s) = f_n(y - s), \tag{5}$$

where $f_n(y - s)$ is the density of the Gaussian noise evaluated at $y - s$. For an unimodal, differentiable a posteriori density, $\hat{s}_{\text{MAP}}$ can be obtained by solving

$$\hat{s}_{\text{MAP}} : \frac{\partial}{\partial s}\left[\ln f_n(y - s) + \ln f(s)\right] = 0. \tag{6}$$

Since $f_n(y - s)$ is known, we readily derive $\hat{s}_{\text{MAP}}$ to be the s for which

$$\frac{s - y}{\sigma^2} + i'(s) = 0, \tag{7}$$

where $i(s) = -\ln f(s)$ is the negative log-density of s, and $i'(s) = \frac{d}{ds}i(s)$, is the score function of s. The resulting function $\hat{s}_{\text{MAP}} = g(y)$ acts as a shrinkage operator on the noisy observation. For certain densities, eq. (7) cannot be solved in closed form. In that case, the following approximation to the MAP estimator may be applied [3]

$$\hat{s} = \text{sign}(y)\ \max(0, |y| - \sigma^2|i'(y)|). \tag{8}$$

3. NIG SHRINKAGE

3.1. The normal inverse Gaussian density

The NIG density is a variance-mean mixture of a Gaussian density with an inverse Gaussian. The stochastic variable s is said to be normal inverse Gaussian if it has a probability density of the form [5]

$$f(s) = \frac{\alpha\delta}{\pi}\frac{\exp[p(s)]}{q(s)}K_1[\alpha q(s)], \tag{9}$$

where $K_1(s)$ is the modified Bessel function of the second kind with index 1, $p(s) = \delta\sqrt{\alpha^2 - \beta^2} + \beta(s - \mu)$, $q(s) = ((s-\mu)^2 + \delta^2)^{1/2}$, $0 \leq |\beta| < \alpha, \delta > 0$ and $-\infty < \mu < \infty$.

The shape of the NIG density is specified by the four-dimensional parameter vector $[\alpha, \beta, \mu, \delta]^T$. The rich parametrization makes the NIG density a suitable model for a variety of unimodal positive kurtotic data. The α-parameter controls the steepness or pointiness of the density, which increases monotonically with increasing α. A large α implies light tails, a small value implies heavy tails. The β-parameter controls the skewness. For $\beta < 0$ the density is skewed to the left, for $\beta > 0$ the density is skewed to the right, while $\beta = 0$ implies a symmetric density around μ, which is a centrality parameter. The δ-parameter is a scale-like parameter.

In [6], Hanssen and Øigård derived a cumulant based estimator for the NIG parameters. By estimating the first four lowest cumulants $\kappa^{(1)}, \kappa^{(2)}, \kappa^{(3)}$ and $\kappa^{(4)}$ from the sample data, and use these to estimate skewness $\hat{\gamma}_3 = \hat{\kappa}^{(3)}/[\hat{\kappa}^{(2)}]^{3/2}$ and normalized kurtosis $\hat{\gamma}_4 = \hat{\kappa}^{(4)}/[\hat{\kappa}^{(2)}]^2$ we form the auxiliary variables

$$\xi = 3\left(\hat{\gamma}_4 - \frac{4}{3}\hat{\gamma}_3^2\right)^{-1}, \qquad \rho = \frac{\hat{\gamma}_3}{3}\sqrt{\xi}. \tag{10}$$

Thereafter, the parameter estimators can easily be shown to be [6]

$$\hat{\delta} = \sqrt{\hat{\kappa}^{(2)}\xi(1-\rho^2)}, \tag{11}$$

$$\hat{\alpha} = \frac{\xi}{\hat{\delta}\sqrt{1-\rho^2}}, \tag{12}$$

$$\hat{\beta} = \hat{\alpha}\rho, \tag{13}$$

$$\hat{\mu} = \hat{\kappa}^{(1)} - \rho\sqrt{\hat{\kappa}^{(2)}\xi}. \tag{14}$$

This estimation technique requires fairly large sample sets to be accurate, and it yields statistically consistent estimators [6].

We now estimate the parameters from the noise free data set, but the parameters can also be estimated from the noisy observations if we know the noise variance σ^2, by subtracting σ^2 from the estimate of $\kappa^{(2)}$. This can be done since the zero mean, Gaussian noise only contributes to the second order cumulant, and in addition it is independent of the signal.

3.2. Shrinkage function

The score function of the NIG density is found to be

$$i'_{NIG}(s) = \frac{\alpha(s-\mu)}{q(s)}\left(\frac{K_0[\alpha q(s)]}{K_1[\alpha q(s)]} + \frac{2}{\alpha q(s)}\right) - \beta. \tag{15}$$

To find $\hat{s}_{\text{MAP}}$ based on the NIG density, $i'_{NIG}(s)$ should be inserted into (7) and the equation solved for s. Unfortunately, the equation is to complex to be solved in closed form. Instead we use the approximate MAP-estimator given in (8).

4. SPARSE CODE SHRINKAGE ALGORITHM

We assume that we have available a large number M of realizations of noise free $(n \times 1)$ random data vectors $\mathbf{x}$. We organize the set of vectors in $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$. These vectors are used to estimate the ICA transform matrix $\mathbf{W}$. This can be done by means of any ICA algorithm. In image denoising, we assume that $\mathbf{X}$ can be obtained from other noise free images of the same category as the image to be denoised. For instance, if we are to denoise an image of a natural scene, we estimate $\mathbf{W}$ based on images of other natural scenes.

The vectors in $\mathbf{X}$, are transformed into a set of sparse $(n \times 1)$ vectors $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_M]$ by the orthogonalized $\mathbf{W}$. Hence $\mathbf{S} = \mathbf{WX}$. The first element of each vector $\mathbf{s}_i$, $i = 1, \ldots, M$, is a realization of independent component no. 1, the second element of independent component no. 2, and so on. The realization of each component is input to the cumulant based NIG parameter estimator, which determines a fit of the NIG density of the noise free component, and we calculate the corresponding shrinkage function. Now, a noisy data vector $\mathbf{x} + \mathbf{v}$ is transformed into its sparse representation $\mathbf{y}$. Every component of $\mathbf{y}$ is denoised according to the NIG shrinkage function associated with each noise free component. The algorithm is summarized as follows [3]

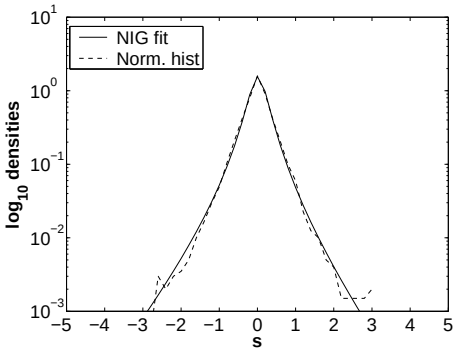
1. Estimate $\mathbf{W}$ from $\mathbf{X}$, and orthogonalize it.
2. Estimate a NIG density for every independent component s_j, $j = 1 \ldots n$, from $\mathbf{S} = \mathbf{WX}$, and find the corresponding shrinkage function g_j.
3. A data vector $\mathbf{x} + \mathbf{v}$ is observed. Make the transformation $\mathbf{y} = \mathbf{W}(\mathbf{x} + \mathbf{v})$.
4. Apply g_j to every element of y_j, to obtain $\hat{s}_j = g_j(y_j)$. Thus $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_n)^T$.
5. Do the inverse transformation $\hat{\mathbf{x}} = \mathbf{W}^T\hat{\mathbf{s}}$.

If the noise variance is unknown it can be estimated by taking the median absolute deviation of the y_i corresponding to the sparsest noise free s_i and divide by 0.6745 [2, 3].

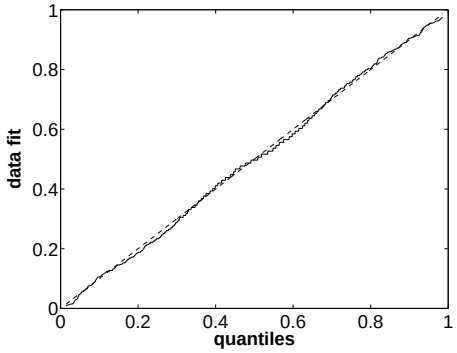
5. APPLICATION TO IMAGE DATA

We have implemented the NIG shrinkage model to denoise images of natural scenes, i.e. images void of any man-made structures. The images are the same as those used in [3]. We express the two-dimensional signals by a one-dimensional column vector by a row-by-row scanning of the image. For computational reasons we can not employ the method directly on the full size images, but rather we extract 10000 (12×12) patches at random from nine different natural scene images. These were ordered in (144×1) vectors.

Several authors have applied ICA to image data, and found that one component represents the local mean image intensity. It was noted in [3] that this component actually does not belong to a sparse density, and that it has a large variance associated with it. Therefore, we ignore this component by subtracting the local mean from each (144×1)

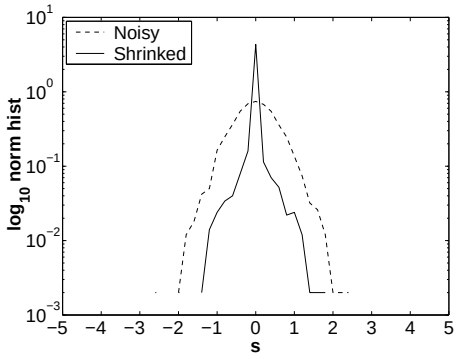


(a) Normalized histogram and fitted NIG density.

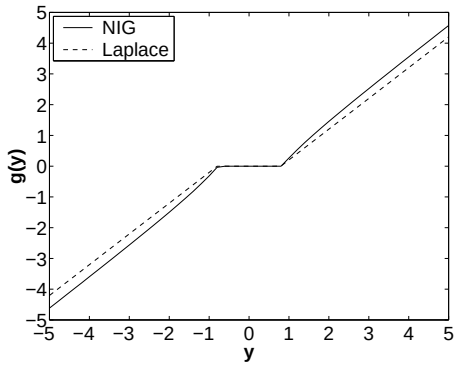


(b) Variance stabilized p-p plot for NIG fit

Fig. 1. Goodness-of-fit for NIG density modeling of noise free IC no. 1.



(a) Normalized histograms, noisy and shrinked



(b) Shrinkage function

Fig. 2. Effect of denoising noisy IC no. 1 by shrinkage function calculated from NIG density fitted to noise free data.

data vector used in the experiment and drop one dimension by principal component analysis (PCA), before proceeding. Thus the dataset consists of vectors of dimension (143×1).

In this manner we had 10000 realizations of noise free data $\mathbf{X}$ to be used in estimating the transform matrix $\mathbf{W}$, the densities of the transformed components, and in calculating shrinkage functions. When estimating $\mathbf{W}$ we used the FastICA [8] algorithm with the hyperbolic tangent nonlinearity.

A separate image was chosen for denoising. In order to control the signal-to-noise ratio (SNR) of the noisy image, the original image was normalized to unit variance before adding noise. In addition the pixels were made zero mean. This was also done for all the images used in estimating the transform.

We slide a (12×12) patch over the noisy image, thus extracting a number of noisy vectors of dimension (143×1). The number of vectors depends on the overlap between the patches. These are ICA transformed, and each vector is denoised. Since the extracted patches overlap we get several suggested values for each denoised image pixel, and we take the result to be the mean of all estimates. In the end, we add the local means to $\hat{\mathbf{x}}$, represent the vectors as (12×12) patches, and order these into a denoised image.

6. RESULTS

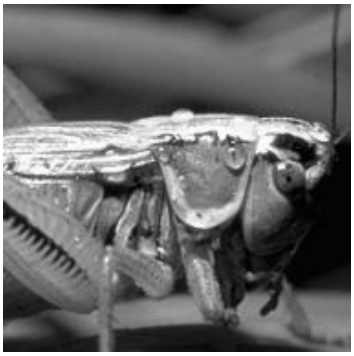
6.1. Fit to data

It is essential that the NIG density actually fits the transformed data. For the purpose of illustrating how close the NIG density models the ICA transformed data, we have used independent component no. 1 (hereafter denoted IC no. 1). The estimated kurtosis of this component was found to be $\hat{\gamma}_4 = 9.03$. The four parameters of the NIG density modeling the underlying probability density function of IC no. 1 were estimated to be

$$\hat{\alpha} = 1.31,\ \hat{\delta} = 0.26,\ \hat{\beta} = -0.08 \text{ and } \hat{\mu} = 0.02. \quad (16)$$

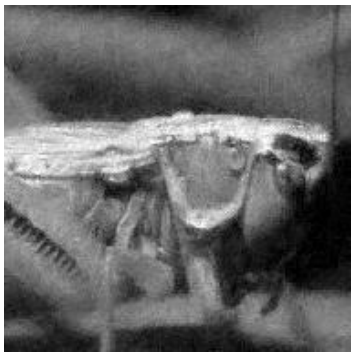
The relatively low value of α indicates a density with rather heavy tails. It has a negligible skewness and it is centered close to origo. The resulting NIG density is shown in a log-plot in figure 1 (a) (solid line) along with a log-plot of the normalized histogram of IC no. 1 (dashed line). The plots only deviate slightly in the tails, giving the first indication of a good data fit.

A common procedure to evaluate the goodness-of-fit of a probability density model is to construct a variance stabi-



(a) Original image

(b) Noisy image, $\sigma = 0.5$



(c) NIG based sparse code shrinkage

(d) Wiener filtered

Fig. 3. Denoising experiment on grashopper image. Restoration of image shown in (b) by, (c) NIG based sparse code shrinkage, and (d) the Wiener filter method.

lized p-p plot [9], which compares inverse sine-transformed uniformly constructed quantiles of the data with inverse sine-transformed quantiles calculated from the exact cumulative distribution of the proposed density. A linear plot indicates a good fit. Given N realizations of a random variable, we construct the uniform quantiles $q_i = (i - 1/2)/N$ and the quantiles $u_i = F(s_i; \alpha, \delta, \beta, \mu)$, $i = 1, \ldots, N$ where $s_1 \leq \cdots \leq s_N$ are ordered samples from the unknown distribution, and F is the exact cumulative distribution of the proposed density. The variance stabilized p-p plot is now defined as the plot of u_i^t against q_i^t, where [9]

$$\begin{aligned} u_i^t &= \frac{2}{\pi} \arcsin(u_i^{1/2}) \\ q_i^t &= \frac{2}{\pi} \arcsin(q_i^{1/2}), \end{aligned} \tag{17}$$

for $i = 1, \ldots, N$. The resulting plot for the NIG model of IC no. 1 is shown in figure 1 (b). It is noted to be very close to linear, indicating that the NIG density of figure 1 (a) does indeed provide an excellent fit to the data.

6.2. Shrinkage function obtained from data

We chose to denoise the "grashopper" image used in [3]. The image was normalized in variance before adding white Gaussian noise with $\sigma = 0.5$, resulting in $SNR = 4$, where SNR is defined as the ratio of the variance in the original image, to the noise variance.

We also used the realizations of noisy IC no. 1 to demonstrate the effect of the shrinking procedure. The dashed curve of figure 2 (a) shows a log-plot of the normalized histogram of noisy IC no. 1. The component is clearly influenced by the Gaussian noise.

Based on the NIG pdf of figure 1 (a), a shrinkage function for the noisy IC no. 1 is calculated. This is shown as the solid non-linearity of figure 2 (b). The horizontal part of the function close to origo corresponds to those components whose non-zero value is thought to be caused by noise only. These are set to zero. The rest of the components are shrinked into values less than the original, but still non-zero. The result of applying the shrinkage function of figure 2 (b) to noisy IC no. 1 is shown in the solid curve of figure 2 (a). This is the normalized histogram of denoised IC no.

1. It is evident that the value of the components has been reduced. The components are concentrated around zero to a much higher degree in the denoised dataset, compared to the noisy dataset.

For comparison, the shrinkage function we would obtain if the data had been modeled by a classical Laplacian density is shown as dashed curve of figure 1 (b). The shrinkage function is in this case given by

$$\hat{s} = \text{sign}(y) \max\left(0, |y| - \frac{\sqrt{2}\sigma^2}{d}\right), \qquad (18)$$

where d is the standard deviation of the density model, easily estimated from the data. In this case the two shrinkage functions have almost identical thresholds. But the large components are shrinked less by the NIG model than by the Laplacian model. The reason for this is that the estimated NIG density has heavier tails than the estimated Laplacian density.

6.3. Denoising a natural image

In figure 3 (a) the "grashopper" image is shown. Figure 3 (b) shows the noisy image, where $\sigma = 0.5$. Figure 3 (c) shows the result of denoising the image of figure 3 (b) by NIG based sparse code shrinkage. Clearly, noise has been effectively reduced, although it still appears a little bit grainy. Compared with the original image, figure 3 (a), it can be seen that some of the contrast is lost, and some blurring is introduced. For comparison, the same image was denoised using a standard Wiener-filter (3×3 mask). The result is shown in figure 3 (d). We note that the Wiener-filter method has not been able to reduce noise as effectively as in NIG based sparse code shrinkage.

7. CONCLUSION

In this paper we introduced the recent normal inverse Gaussian density as a model for sparsely coded data. The NIG density is a flexible, four-parameter density, highly suitable for modeling possibly skewed super-Gaussian data. In the symmetric case it can model data having kurtosis of all positive values.

The NIG-density is an alternative pdf to those used e.g. in [3] for modeling sparsely coded data. The parameter estimates for the densities in [3] require an estimate of the peak value of the proposed density, obtained via a non-parametric kernel method. This may be problematic since the estimate relies heavily on the kernel bandwidth. In the NIG case, very fast and simple cumulant based parameter estimators exist, which for fairly large datasets yield accurate results. The parameters can moreover be estimated based on the noisy observation.

We obtained sparsely coded image data by applying ICA to the images, and demonstrated that the NIG density is capable of providing a very good fit to the ICA transformed data.

A NIG-based maximum a posteriori estimator technique was developed. It acts as a shrinkage operator on the noisy ICA transformed data. We demonstrated the NIG-based sparse code shrinkage technique on the "grashopper" image contaminated by Gaussian noise. The technique proved effective in reducing the noise, comparable to the result obtained in [3].

8. REFERENCES

[1] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," *3rd IEEE Int'l Conf. on Image Processing, September, Lausanne, Switzerland*, 1996.

[2] D. L. Donoho and I. M. Johnstone, "Wavelet shrinkage: Asymptopia?," *J. R. Statist. Soc. B*, vol. 57, no. 2, pp. 301–369, 1995.

[3] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[4] P. Z. Peebles, *Probability, random variables, and random signal principles*, McGraw-Hill, Third edition, 1993.

[5] O. E. Bandorff-Nielsen, "Normal inverse Gaussian distributions and stochastic volatility modeling," *Scand. J. Statist.*, vol. 24, pp. 1–13, 1997.

[6] A. Hanssen and T. A. Øigård, "The normal inverse Gaussian distribution as a flexible model for heavy tailed processes," *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, June 3-6, Baltimore, Maryland, USA*, 2001.

[7] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[8] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, pp. 1483–1492, 1997.

[9] J. R. Michael, "The stabilized probability plot," *Biometrika*, vol. 70, pp. 11–17, 1983.