

A model selection approach to signal denoising using Kullback's symmetric divergence

Maïza Bekara, Luc Knockaert, Abd-Krim Seghouane, Gilles Fleury

▶ **To cite this version:**

Maïza Bekara, Luc Knockaert, Abd-Krim Seghouane, Gilles Fleury. A model selection approach to signal denoising using Kullback's symmetric divergence. Signal Processing, 2006, Vol. 86, pp. 1400 -1409. 10.1016/j.sigpro.2005.03.023 . hal-00260925

HAL Id: hal-00260925

https://centralesupelec.hal.science/hal-00260925

Submitted on 5 Mar 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A model selection approach to signal denoising using Kullback's symmetric divergence

Maïza Bekara[a,*], Luc Knockaert[b], Abd-Krim Seghouane[a], Gilles Fleury[a]

[a]*École Supérieure d'Électricité—Servicedes Mésures, 3 rue Joliot Curie 91192 Gif-sur-Yvette Cedex, France*
[b]*IMEC-INTEC-UGENT, St. Pietersnieuwstraat 41, B-900, Gent Belgium.*

Abstract

We consider the determination of a soft/hard coefficients threshold for signal recovery embedded in additive Gaussian noise. This is closely related to the problem of variable selection in linear regression. Viewing the denoising problem as a model selection one, we propose a new *information theoretical* model selection approach to signal denoising. We first construct a statistical model for the unknown signal and then try to find the best approximating model (corresponding to the denoised signal) from a set of candidates. We adopt the *Kullback's symmetric divergence* as a measure of similarity between the unknown model and the candidate model. The best approximating model is the one that minimizes an *unbiased estimator* of this divergence. The advantage of a denoising method based on model selection over classical thresholding approaches, resides in the fact that the threshold is determined automatically without the need to estimate the noise variance. The proposed denoising method, called KIC*c*-denoising (Kullback Information Criterion corrected) is compared with cross validation (CV), minimum description length (MDL) and the classical methods *Sure*Shrink and *Visu*Shrink via a simulation study based on three different type of signals: chirp, seismic and piecewise polynomial.

Keywords: Signal denoising; Model selection; Information criterion

1. Introduction

In this paper, we consider the problem of recovering an unknown signal from sampled noisy data. Given a set of observed data $\mathbf{g}_n = (g_1, g_2, \ldots, g_n)^\top$, which is assumed to be generated from the model:

$$g_i = f(t_i) + \varepsilon_i, \tag{1}$$

where f is the unknown function to be estimated sampled at the instants t_i, ε_i are noise samples assumed to be i.i.d Gaussian random variables with zero mean and variance σ_0^2 and let $\mathbf{f}_n = (f(t_1), f(t_2), \ldots, f(t_n))^\top$. Our objective is to obtain $\hat{\mathbf{f}}_n$, an estimate of $\mathbf{f}_n$ based on the observations $\mathbf{g}_n$.

Classically, this problem has been solved by linear processing through dynamic filtering. However, in many interesting cases, linear methods fail to give satisfactory results. Effective denoising methods are frequently based on some nonlinear processing, which consists of a cascade of three mappings:

1. linear mapping W, that transforms the data by projecting it on a basis, usually taking to be an orthogonal one (wavelet, fourier, spline, ...). The

*Corresponding author.
E-mail addresses: Maiza.Bekara@supelec.fr (M. Bekara), knokaert@intec.rug.ac.be (L. Knockaert).

resulting transformed data $\mathbf{y}_n = W\mathbf{g}_n$, is called the *coefficients*.

2. nonlinear mapping $F(.)$, used to map the coefficients $\mathbf{y}_n$ to obtain a new vector of coefficients, i.e., $\hat{\mathbf{c}}_n = F(\mathbf{y}_n)$. Two known mapping functions, which are called *thresholding functions* are defined as:

   Hard-thresholding: $F_h(y, \lambda) = yI(|y| > \lambda)$,

   Soft-thresholding:

   $$F_s(y, \lambda) = \text{sgn}(|y| - \lambda)(|y| - \lambda)I(|y| > \lambda),$$

   where $I(.)$ and sgn(.) are indicator and sign functions, respectively. λ is a positive real, known as the *threshold*.
3. Linear mapping: an inverse transformation, W^{-1} is applied to the denoised signal, i.e., $\hat{\mathbf{f}}_n = W^{\top}\hat{\mathbf{c}}_n$.

Under the above scheme, the problem of denoising is reduced to that of finding an appropriate value of λ. This choice is very crucial and may lead to oversmoothing (if λ is too large) or undersmoothing (if λ is too small). Donoho and Johnstone proposed simple and explicit expressions to compute the optimal threshold. *Visu*Shrink [1] and *Sure*Shrink [2] are two methods widely used in signal processing. However, their implementation needs to compute an estimate of the noise variance and therefore making their performance highly dependent on the quality of this estimation.

Since the thresholding is applied to the coefficients, it sounds to think that the problem of finding an optimal value for λ is not really a continuous optimization problem as considered in [1,2], but a discrete one. In this case, the denoising problem is viewed as that of finding the most significant projection bases (coefficients) rather than finding the numerical value of the threshold. This is equivalent to a model selection formalism for linear regression which is aimed to find the best *predictor* to explain a dependent variable. Here, the number of all possible predictor (bases) is equal to the number of data points n. It is required to select the k most significant bases out of the n possible, therefore setting the rest $(n - k)$ bases to zero, and this is equivalent to thresholding. Viewing the denoising problem in this way is of great interest because it leads to an automatic determination of the threshold without the need to directly estimate the noise variance.

In this paper, we extend the class of denoising methods based on model selection. We propose an *information theoretical* approach to signal denoising based on Kullback's symmetric divergence. We first assume a statistical framework for the unknown generating model and then we try to find the best approximating model within a nested parametric classes of models of increasing complexity k. We adopt the *Kullback symmetric divergence*, known also as the J-divergence [3] as a measure of similarity between the true unknown model (ideal signal) and the approximating model (denoised signal). We then select the model that minimizes an exactly unbiased estimate of this divergence.

The rest of the paper is organized as follows. Section 2 reviews the main model selection criteria for signal denoising. Section 3 presents the proposed denoising method which we call the KIC_c-denoising. In Section 4 we present our simulation results along with some discussions and concluding remarks.

2. Denoising and model selection

The model selection approach to signal denoising is not a new issue and has already been described in the literature. We roughly discuss two major approaches:

- The principle of minimum description length (MDL) introduced by Rissanen [4] and rooted in *coding theory* was the first model selection approach to be applied for signal denoising [5,6] and image enhancement [7]. The MDL principle suggests that the best model among a given collection of candidate models is the one that gives the shortest description of the data and the model itself. As a consequence, the noise is defined as the part of the data that cannot be compressed by the model, while the rest defines the meaningful information-bearing signal. Therefore, the MDL principle simultaneously provides the best *denoising* and *compression* of the signal. A comprehensive paper on MDL-denoising is given by Rissanen [8].
- Vapnik–Chervonenkis (VC) theory is another principle for function estimation which emerged from *learning theory* [9]. VC-theory provides an analytical upper bound of the prediction risk, which can be used for model selection. The strength of VC-theory resides in the fact that the proposed bound is obtained for finite samples, therefore avoiding asymptotic approximations and it is distribution independent, i.e., no need to have an assumption about the noise distribution.

However, the so called VC-denoising suffers from some drawbacks. To compute the bound it is required to provide an accurate estimate of the candidate model's VC-dimension. Generally it is very hard to obtain a closed form expression of the VC-dimension and hence it is roughly approximated by the number of free parameters in the model. Others implementation difficulties are encountered when VC-denoising is applied; some constants in the definition of the upper bound are set without formal arguments and the proposed model's structures are not theoretically justified [10]. Moreover, the performance of VC-denoising shows interesting results only if the assessment is based on "best worse-case" of the quadratic risk,[1] whereas the classical assessment is based on "best average-case". For the above reasons we will consider only MDL denoising in the simulations examples.

The approach we propose is different from the previous ones in the following sense. First, as compared with VC-denoising, it is based on an unbiased estimate of the risk rather then an upper bound. Secondly, it avoids the dual objective of compression and denoising achieved by the MDL principle and concentrates only on denoising which is the objective of interest.[2]

3. KIC$_c$-denoising

Let us introduce the binary vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)^\top \in \{0,1\}^n$ as an index of 2^n possible models $\mathcal{M}_\gamma$. The elements of γ specify which regressor is included in the model, hence we have:

$$c_i \neq 0 \quad \gamma_i = 1$$

$$c_i = 0 \quad \gamma_i = 0$$

Under the model $\mathcal{M}_\gamma$, the vector of data $\mathbf{g}_n$ exhibits the Gaussian distribution

$$p(\mathbf{g}|\gamma, \mathbf{c}_\gamma, \sigma_\gamma^2, \mathcal{M}_\gamma) = \left(2\pi\sigma_\gamma^2\right)^{-n/2} \exp\{-\|\mathbf{g} - W^\top \mathbf{c}_\gamma\|^2/2\sigma_\gamma^2\}, \quad (2)$$

where $\mathbf{c}_\gamma$ is the vector of coefficients which are nonzero at the locations indexed by γ and σ_γ^2 is a parameter which corresponds to the noise variance.

Let $\mathbb{M}_k = \{\mathcal{M}_\gamma | \sum_{i=1}^n \gamma_i \leqslant k\}$ denote the family of models with *at most* k regressors. The $\mathbb{M}_k$'s have a form of nested structure of models, i.e., $\mathbb{M}_k \subset \mathbb{M}_{k+1}$. In order to reduce the number of candidate models, we try to find the best model within a given family of complexity k. For that purpose we consider the model $m_k \in \mathbb{M}_k$ that maximizes the log likelihood over its class, i.e.,

$$m_k = \arg \max_{\mathcal{M}_\gamma \in \mathbb{M}_k} \log p(\mathbf{g}_n|\gamma, \mathbf{c}_\gamma, \sigma_\gamma^2, \mathcal{M}_\gamma). \quad (3)$$

Let $\hat{\mathbf{c}}_k$ denote the maximum likelihood estimation (MLE) of the coefficients for the model m_k, formally obtained by:

$$\begin{aligned} \hat{\mathbf{c}}_k &= \arg\min_{\mathbf{c}_\gamma} \{\|\mathbf{g}_n - W^\top \mathbf{c}_\gamma\|^2\} \\ &= \arg\min_{\mathbf{c}_\gamma} \{\|\mathbf{y}_n\|^2 - \|\mathbf{c}_\gamma\|^2\}. \end{aligned}$$

Clearly, in the family of models $\mathbb{M}_k$, the best approximating model is the one obtained by taking the k largest coefficients in absolute value. We denote this vector by $\hat{\mathbf{c}}_k$. Now instead of selecting the best model from a set of 2^n candidates, we will have to compare only n models. Similarly, if σ_k^2 denotes the MLE of the noise variance under the model m_k. It will be equal to:

$$\sigma_k^2 = \{\|\mathbf{y}_n\|^2 - \|\hat{\mathbf{c}}_k\|^2\}/n. \quad (4)$$

Let us assume that the true model has a complexity $k_0 \in [0, n]$. This is a strong assumption, but it is usually employed in developing model selection criterion. Although the requirement may seem strong, a criterion developed under this assumption often achieves its intended objectives, even when the assumption is violated (see [11, pp. 20–22]). This assumption is in fact equivalent to say that the ideal signal has at most k_0 nonzero wavelet coefficients and the rest $(n - k_0)$ are zeros. Every test signal considered in our simulations fulfill this assumption.

There exist different measures of similarity between probability density functions. The most frequently used measure is probably the *Kullback–Leibler divergence*, also known as the I-divergence or simply Kullback directed divergence. The Kullback directed divergence between two statistical

[1] Best worse-case is indicated by lower 75% and 95% marks of the mean squared error when Monte Carlo simulations are performed.

[2] In fact denoising results in compression and vice versa. Therefore, all the denoising methods based on thresholding results in data compression. However, for the MDL case, the compression objective is put forward rather than being a result.

models $p(\mathbf{x}|\boldsymbol{\theta}_0)$ and $p(\mathbf{x}|\boldsymbol{\theta}_k)$ is defined as:

$$2I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) = \mathbb{E}_0\left\{2 \log \frac{p(\mathbf{x}|\boldsymbol{\theta}_0)}{p(\mathbf{x}|\boldsymbol{\theta}_k)}\right\} = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0), \tag{5}$$

where

$$d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \mathbb{E}_i\{p(\mathbf{x}|\boldsymbol{\theta}_j)\},$$

and the expectation $\mathbb{E}_i\{.\}$ is with respect to the distribution $p(\mathbf{x}|\boldsymbol{\theta}_i)$.

The Kullback directed divergence is an asymmetric measure, which means that an alternative directed divergence can be obtained by reversing the roles of the two models in (5). A new measure of dissimilarity can be obtained by the sum of the two directed divergences. This new measure is known as the Kullback's symmetric divergence, or J-divergence [3]. Since the J-divergence combines information about the model's dissimilarity through two distinct measures, it functions as a gauge of model disparity, which is arguably more sensitive than either of its individual component. The Kullback symmetric divergence is defined as

$$\begin{aligned} 2J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) &= 2I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) + 2I(\boldsymbol{\theta}_k, \boldsymbol{\theta}_0) \\ &= d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) + d(\boldsymbol{\theta}_k, \boldsymbol{\theta}_0) \\ &\quad - d(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0). \end{aligned} \tag{6}$$

Dropping $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ since it does not depend on k, the quantity

$$K(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) + d(\boldsymbol{\theta}_k, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k), \tag{7}$$

is a suitable substitute measure for $2J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k)$. If we denote $\hat{\boldsymbol{\theta}}_k$ the MLE of $\boldsymbol{\theta}_k$ based on a vector of observed data $\mathbf{x}_n$, $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_k)$ would provide a suitable measure of the discrepancy between the two models. Yet evaluating $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_k)$ is not possible since it requires the knowledge of the true model. It has been shown [12] that for over-parameterized or exactly specified models (the unknown model belongs to the class of candidate), and under the hypothesis of linear regression models with i.i.d Gaussian noise, an exactly unbiased estimator of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_k)$ denoted by KIC_c is given by:

$$\mathrm{KIC}_c(k) = -2 \log p(\mathbf{x}_n|\hat{\boldsymbol{\theta}}_k) + 2\frac{(k+1)n}{n-k-2} - n\psi\left(\frac{n-k}{2}\right) + n \log\left(\frac{n}{2}\right), \tag{8}$$

where $\psi(.)$ is the *digamma* function [13]. Under this assumption the goodness of fit term determined by the $-\log$ likelihood is equal to

$$-2 \log p(\mathbf{x}_n|\hat{\boldsymbol{\theta}}_k) = n \log(\sigma_k^2) + n \log(2\pi) + n.$$

Using the result in (4) and dropping the constant terms from (8), the criterion reduces to

$$\mathrm{KIC}_c(k) = n \log\left(\frac{\|\mathbf{y}_n\|^2 - \|\hat{\mathbf{c}}_k\|^2}{n}\right) + 2\frac{(k+1)n}{n-k-2} - n\psi\left(\frac{n-k}{2}\right). \tag{9}$$

The KIC_c has been implemented to solve divers model order selection problems including polynomials, AR, ARMA models [14] and source number estimation in array processing [15]. It is worth to mention that the assumptions made above, mainly the one that assumes a Gaussian noise, are made only in the derivation of the criterion. When these assumptions are not fulfilled, (8) and (9) are just *approximately* unbiased estimators instead of exactly unbiased. It is then possible to study the performance of KIC_c without regard to the assumptions underlying the derivation.[3] However, this issue needs further analytical development and extensive empirical simulations which are beyond the scope of this paper.

Using the above criterion, we propose the KIC_c-denoising algorithm as follows:

KIC_c-denoising

1. Obtain the noisy w coefficients $\mathbf{y}_n = W\mathbf{y}_n$.
2. Order the absolute value of the $\mathbf{y}_n$ in a decreasing order, i.e., $|y|_1 \geqslant |y|_2 \geqslant \cdots > |y|_n$
3. Find the integer $k_{KICc} \in [1, n]$ that minimizes (9). Choose the threshold $\lambda_{opt} = |y|_{k_{KICc+1}}$
4. Apply the soft-thresholding to obtain $\hat{\mathbf{y}}_n = F_s(\mathbf{y}_n, \lambda_{opt})$ and get the denoised signal by inverse transform $\hat{\mathbf{f}}_n = W^{-1}\hat{\mathbf{y}}_n$.

A similar MDL-denoising algorithm [8] can be obtained by taking

$$\mathrm{MDL}(k) = (n-k) \log\left(\frac{\|\mathbf{y}_n\|^2 - \|\hat{\mathbf{c}}_k\|^2}{n-k}\right) + k \log\left(\frac{\|\hat{\mathbf{c}}_k\|^2}{k}\right) - \log\left(\frac{k}{n-k}\right) \tag{10}$$

[3] We have studied the performance of KIC_c in the linear regression case when the assumption that the true model is correctly specified or overfitted is not verified in [16].

instead of (8). A hard thresholding can also be applied by changing the thresholding function in Step 4.

The choice of the orthogonal projection, W depends on the type of the signal in question. For a wide class of interesting signals which include nonstationary signals or signals with discontinuities, projection based *wavelet* decomposition is the most suitable one. In this work we have used this type of decomposition. The advantage of using wavelets to approximate an arbitrary function resides in their ability to spatially adapt to salient features of the function leading to a parsimonious representation. A comprehensive discussion about the theory and applications of wavelets can be found in the book written by Mallat [17].

4. Simulation

In this section, we compare the performance of the proposed KIC_c-denoising method with MDL-denoising, cross validation (CV) [18], *Sure*Shrink (hard thresholding) and *Visu*Shrink (soft thresholding). All simulations are carried out using the WaveLab package developed at Stanford University.[4] We would like to mention that *Sure*Shrink and *Visu*Shrink are *adaptive* thresholding methods, i.e., an appropriate threshold is computed and applied for each resolution. On the other hand, CV, MDL and KIC_c are *global* thresholding strategies, i.e., a single threshold is applied to all the resolutions. Throughout all the simulations, we consider a Daubechies type mother wavelet with $N = 6$ and a lower resolution cutoff $L = 3$.

We consider three types of signals in this paper. The first signal is a *chirp* signal, which is an example of signals with a complex frequency profile. In addition, this type of signals is often encountered in radar and sonar signal processing. The second signal is a seismic signal, as an example of a complex time profile signal. The last signal is a piecewise polynomial signal with a discontinuity which is an example of a smooth signal with a transition, often encountered in edge detection. All simulations are implemented for two different noise levels, low and high respectively. Signals are made noisy by adding i.i.d Gaussian random variables with zero mean and variance σ_0^2. For each combination of signal type and noise level, 1000 different realizations of the noisy signal were generated. For each realization a denoised signal $\hat{\mathbf{f}}_n$ is obtained by using the different competing methods and the mean square error (MSE) is computed as $\|\mathbf{f} - \hat{\mathbf{f}}_n\|^2/n$. The sampling distribution of $\log(MSE)$ is displayed using boxplots and used as performance assessment to compare the different denoising methods.

4.1. Example 1: chirp signal

The chirp signal is defined as

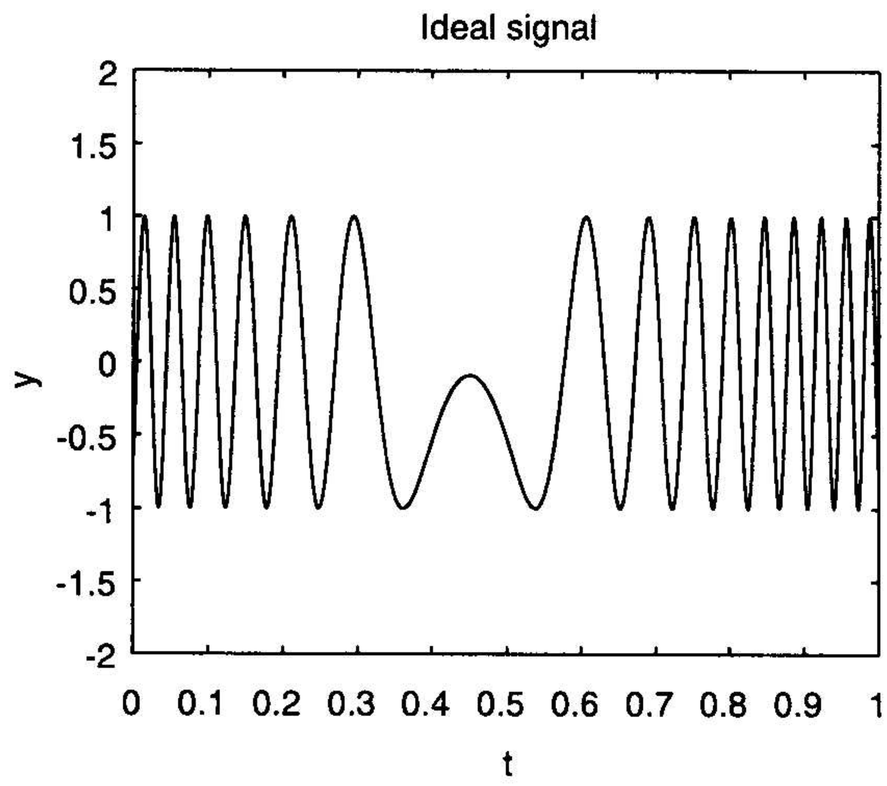
$$f_1(t) = \sin\left(40\pi\left(1.5t^2 - 1.36t + 0.68\right)\right), \quad t \in [0, 1].$$

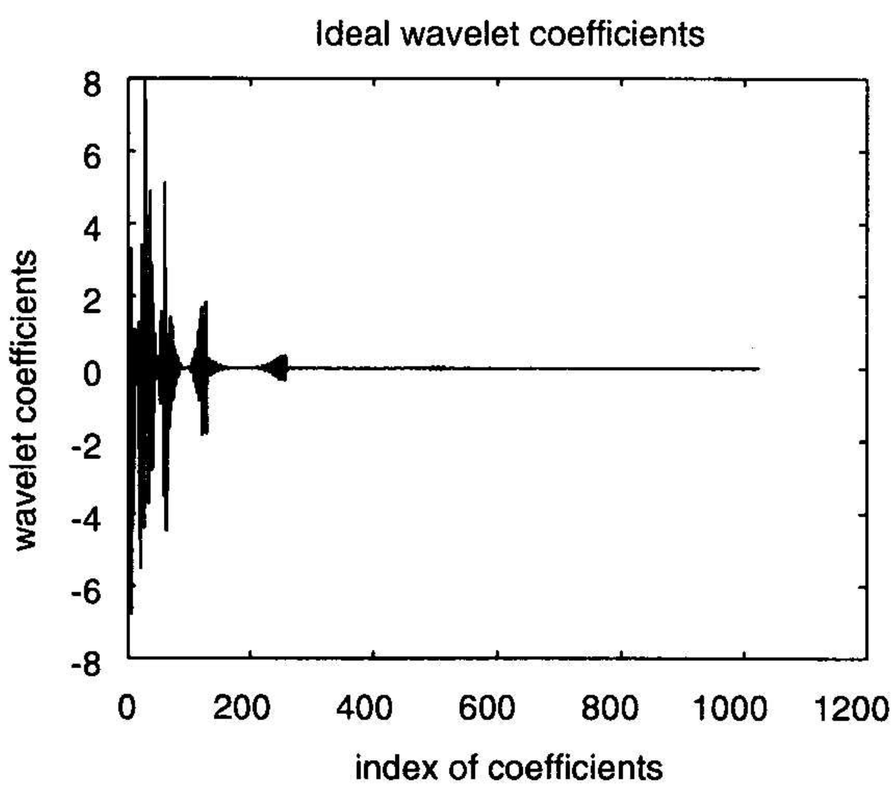
Two noise levels are considered, $\sigma_0^2 = 0.5$ and $\sigma_0^2 = 0.05$. The signal $f_1(t)$ and its wavelet coefficients along with their noisy versions are shown in Fig. 1. Fig. 2 gives boxplots of $\log(MSE)$ for each of the five denoising methods. Clearly KIC_c performed the best for both noise levels, closely followed by MDL. On the other hand, *Visu* was far behind all the remaining methods. This is mainly due to its tendency to oversmooth the noisy signal by setting high threshold. We may notice also the *robust* behavior of *KICc* in term of lower variability of $\log(MSE)$.

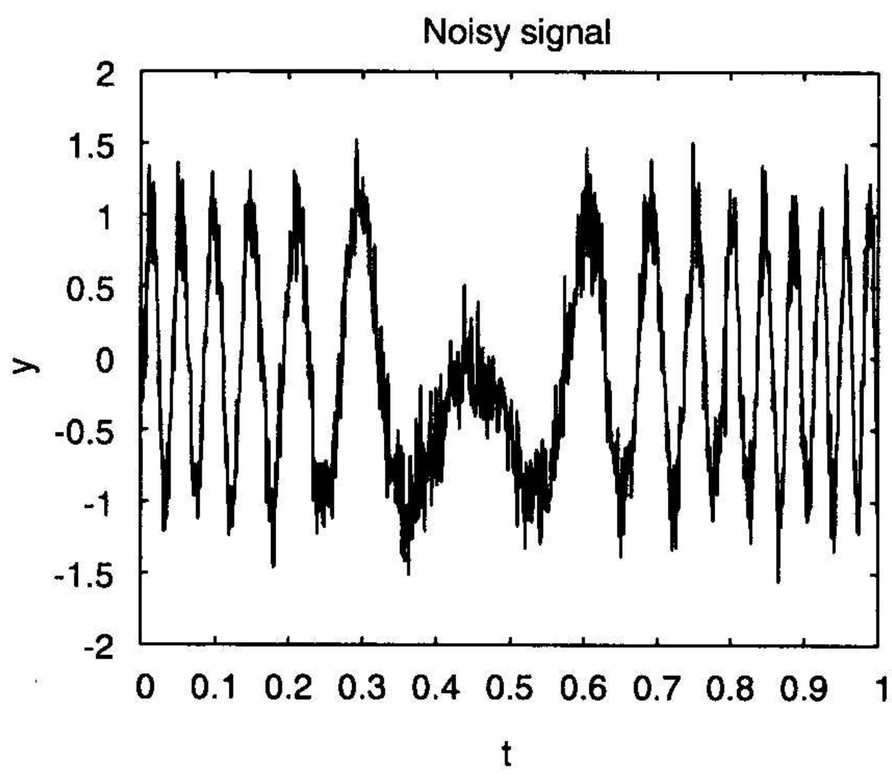
4.2. Example 2: seismic signal

One of the key problems in seismology is to derive information about the structure and physical properties of the earth medium from the analysis of seismic records. This task is complicated by the fact that the seismic signals emitted by the source are weakened by geometric spreading and attenuation. Moreover, they are also distorted by ambient seismic noise. Therefore, one of the main issues in applied seismology is to ensure high SNR or, when conditions are bad, to improve it by suitable ways of data acquisition and processing (denoising). In this example we target a 1-D seismic signal that consists of 1024 time samples. The noise variance σ_0^2 is chosen proportional to the maximum amplitude of the ideal signal, i.e., $\sigma_0^2 = \alpha \max(|y|)$, with $\alpha = \{10^{-2}, 10^{-3}\}$. Fig. 3 shows the seismic signal and its wavelet transform along with their noisy versions. Fig. 4 presents boxplots of the $\log(MSE)$ for each of the denoising methods. Here again, the conclusions obtained for the chirp signal can be

[4] This package is freely available on-line at http://www-stat.stanford.edu/~wavelab/







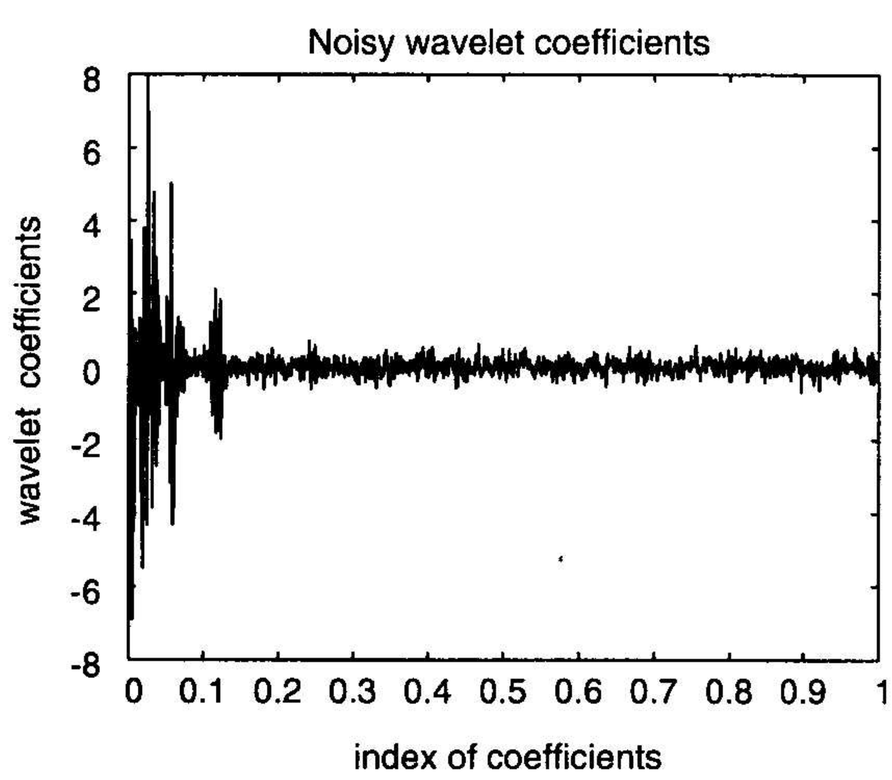
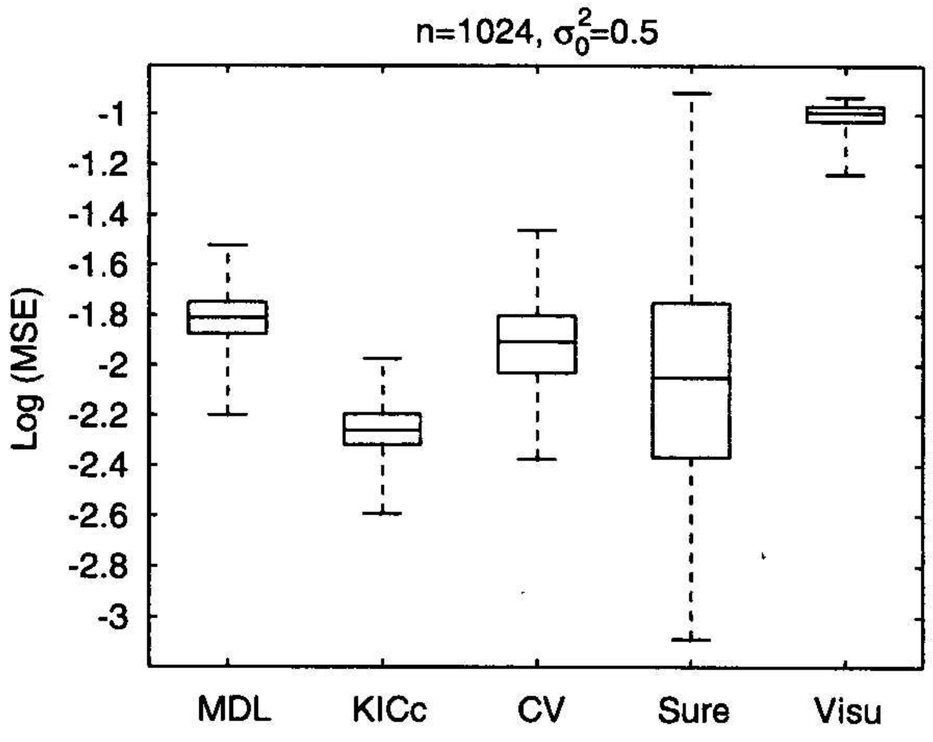


Fig. 1. Chirp signal and its wavelet coefficients, $n = 1024$ and $\sigma_0^2 = 0.05$.



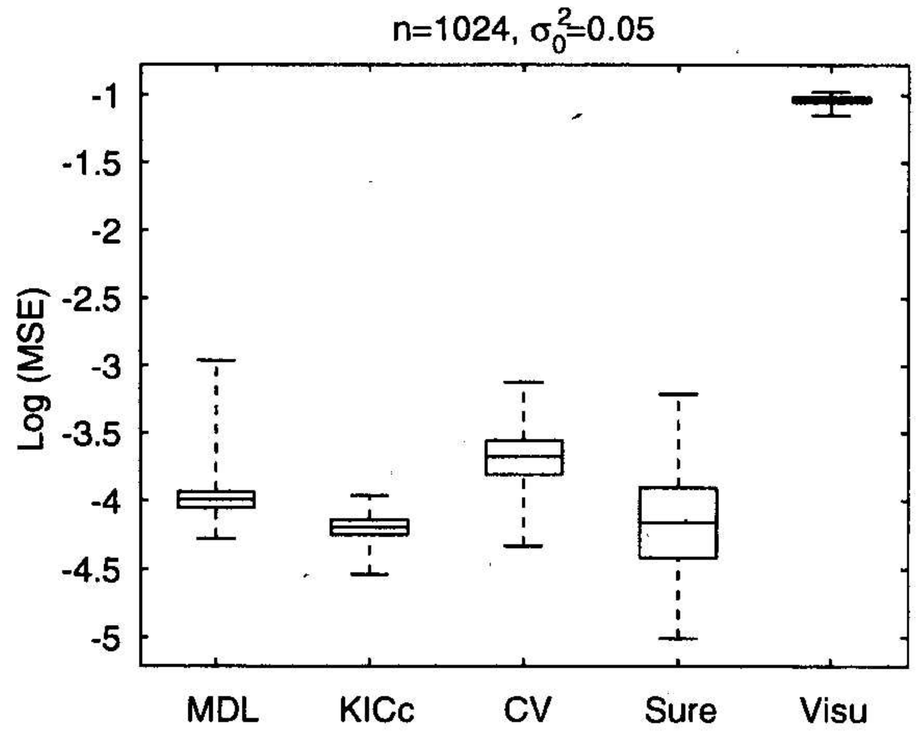


Fig. 2. Boxplot of $\log(MSE)$ of different denoising methods for the chirp signal.

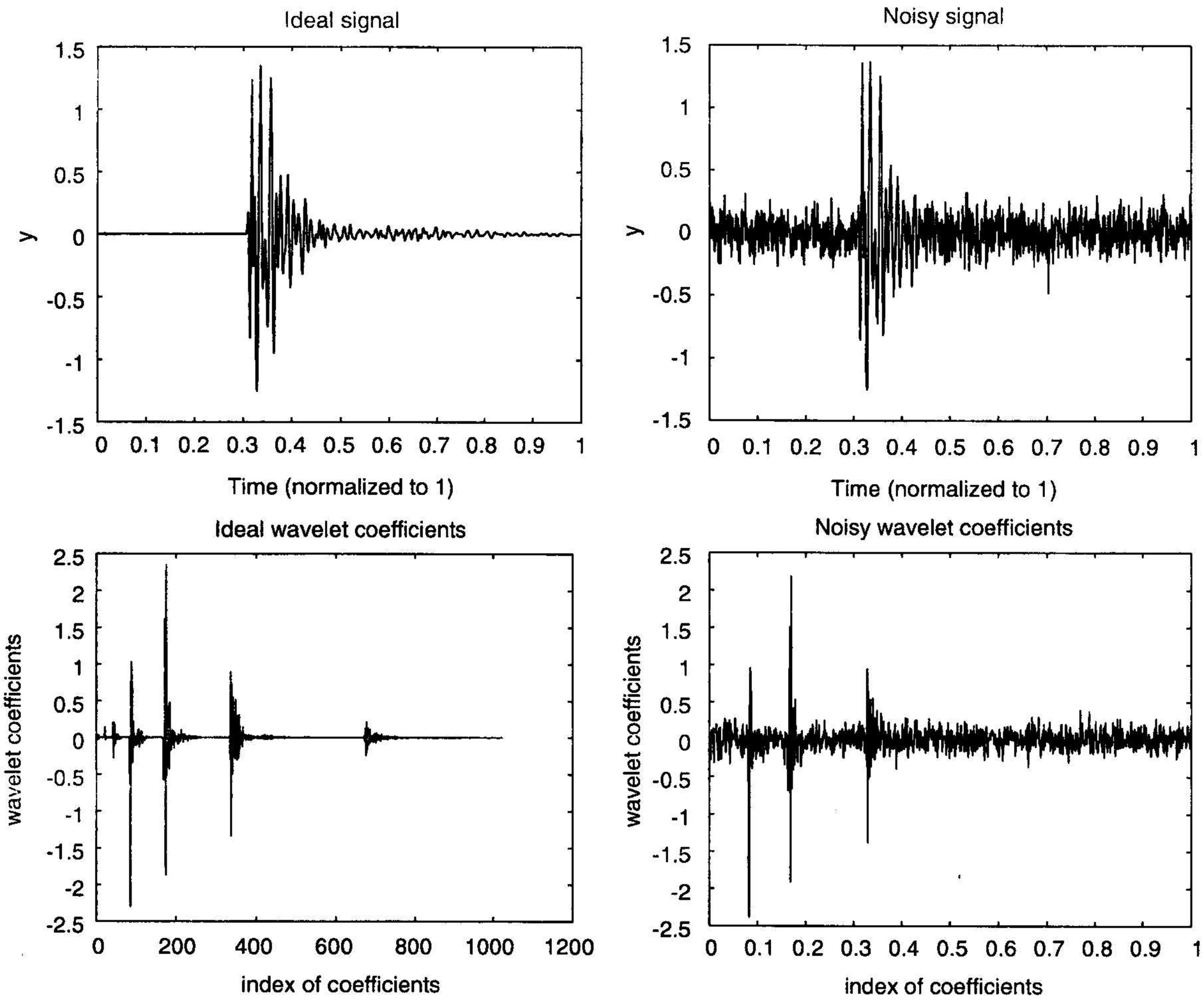


Fig. 3. Seismic signal and its wavelet coefficients, $n = 1024$ and $\alpha = 10^{-2}$.

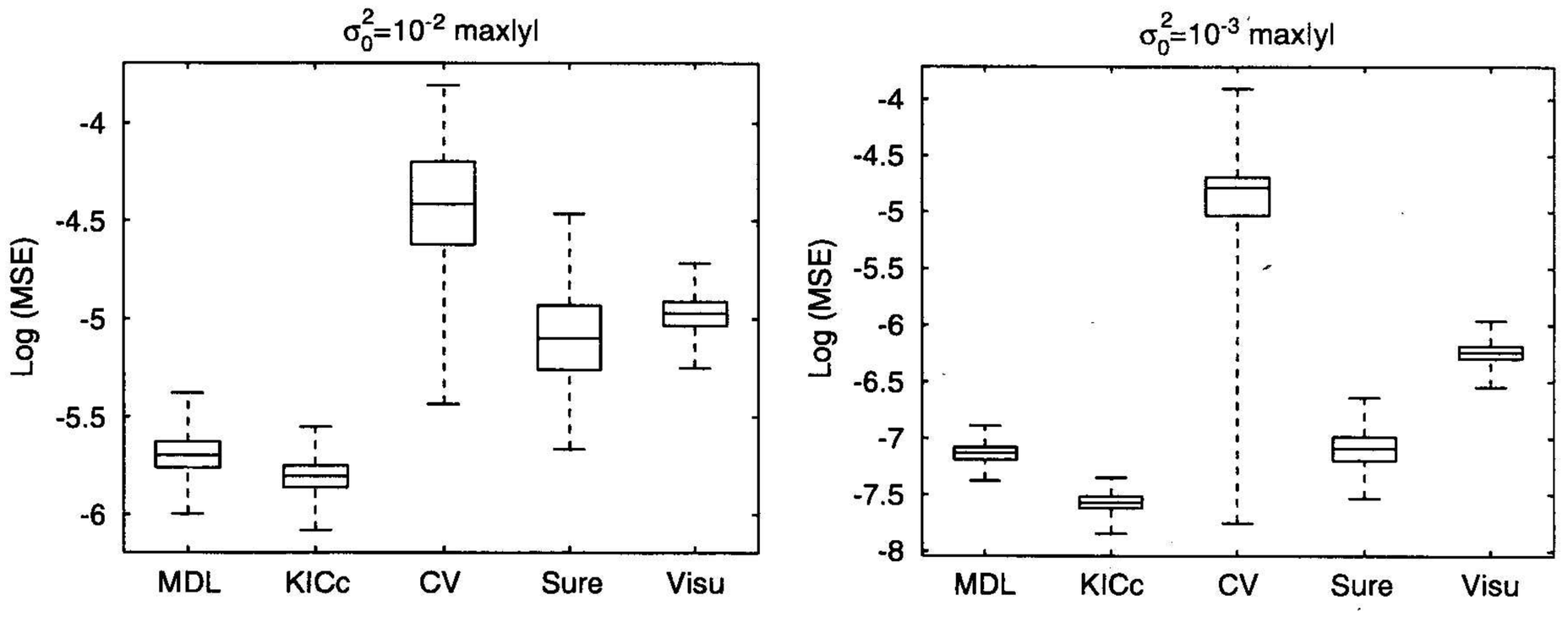


Fig. 4. Boxplot of $\log(\hat{MSE})$ of different denoising methods for the seismic signal.

generalized for the case of seismic signal. The optimality of denoising methods based on model selection over the other methods is clear. A close look at the wavelet coefficients of the ideal signal (the chirp and more clearly the seismic signal) shows a large disparity of the coefficients over

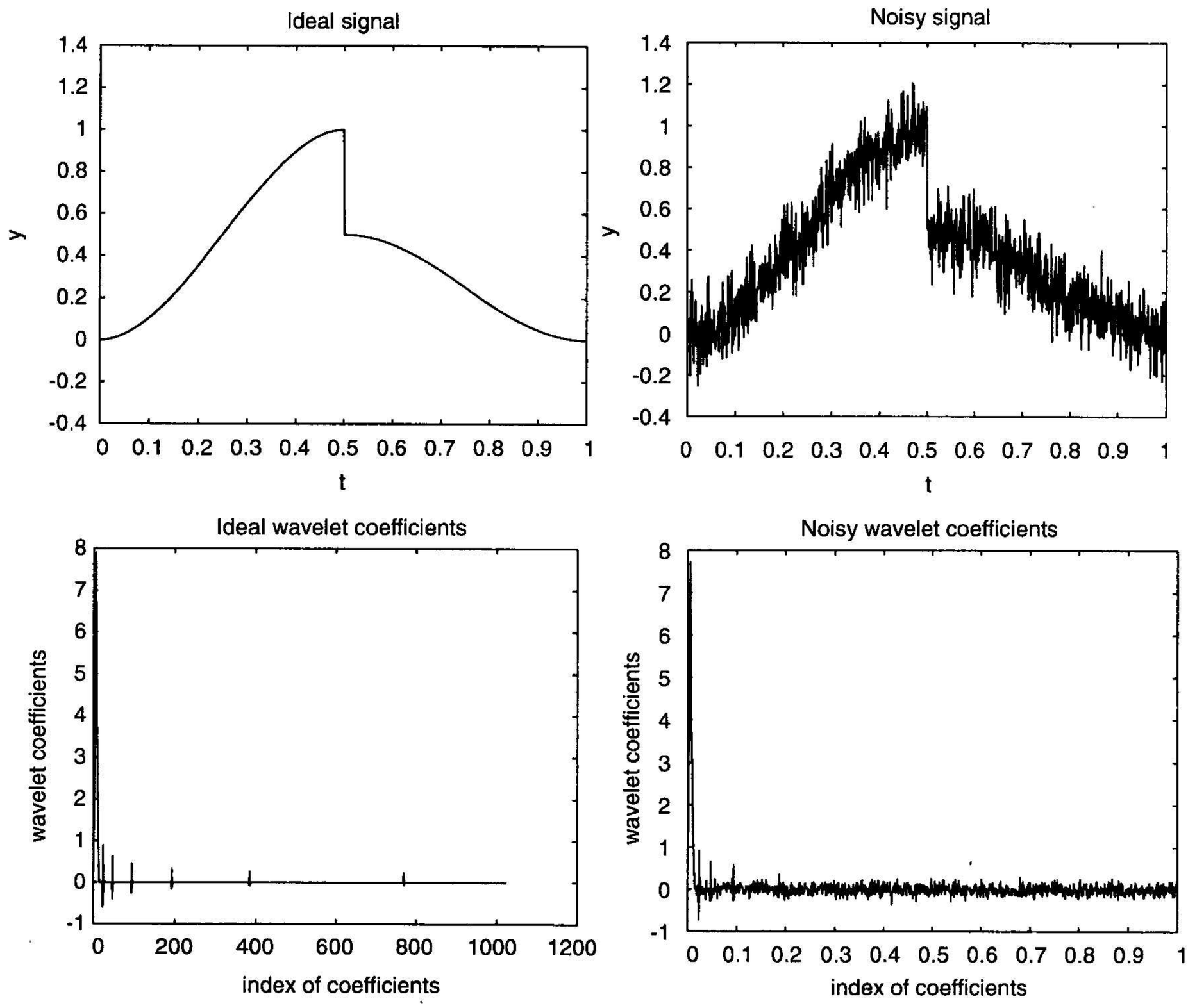


Fig. 5. Piecewise polynomial signal and its wavelet coefficients, $n = 1024$ and $\sigma_0^2 = 10^{-2}$.

all the subbands. This implies that some of the high frequency signal components present in the finest or most detailed subbands is going to be confused with noise. As a consequence, *Visu* tends to overestimate the noise variance, leading to a high threshold and therefore resulting in oversmoothing of the data. On the other hand, *Sure*, which is already known to undersmooth the signal, will excessively undersmooth it, misled by the noise-like components of the signal. However, KIC_c and MDL find a balance in a fully automatic way.

4.3. *Example 3: piecewise polynomial*

The piecewise polynomial with a discontinuity is a well-known academic example presented in [19]. It is defined over the interval [0,1] as

$$f_2(t) = \begin{cases} 4t^2(3-4t) & t \in [0.00, 0.50], \\ \frac{3}{4}t(4t^2-10t+7) - \frac{3}{2} & t \in]0.50, .75], \\ \frac{16}{3}t(t-1)^2 & t \in]0.75, 1.00]. \end{cases}$$

The data consist of 1024 uniform samples over [0,1]. The function $f_2(t)$ and its wavelet coefficients together with their noisy versions are shown in Fig. 5. By examining the wavelet coefficients of the ideal signal, we observe that the signal has effectively very few nonzero coefficients, with many of them caused by the presence of the discontinuity. Now, any thresholding method that sets most of these coefficients equal to zero, will effectively

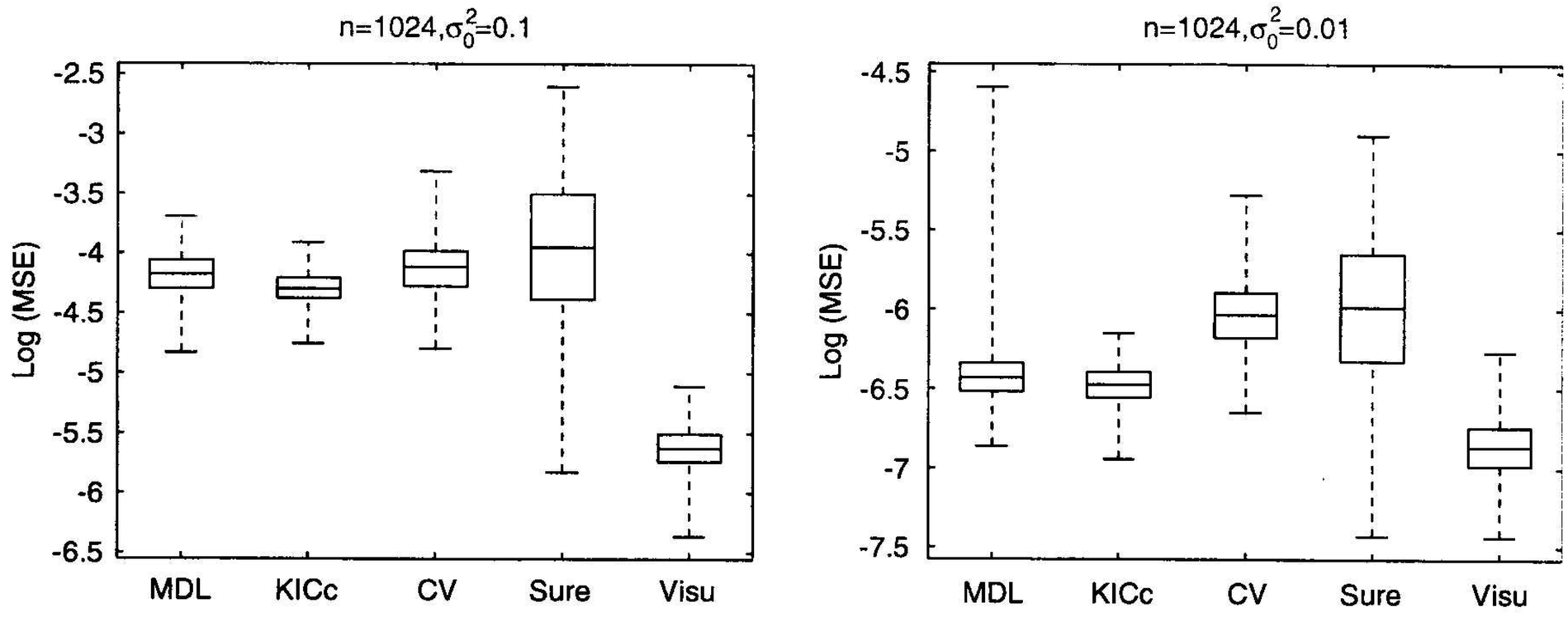


Fig. 6. Boxplot of log(*MSE*) of different denoising methods for the piecewise polynomial signal.

recover the signal at the cost of smoothing the discontinuity. This can be shown by observing the boxplots in Fig. 6. Clearly *Visu*, which was blamed to set a high threshold and have a tendency to oversmooth the data performed the best. Yet KIC_c came in the second place followed by MDL. Again denoising methods based on model selection yield good results in comparison with CV and *Sure*.

5. Conclusion

In this paper we have introduced a new information theoretical criterion for signal denoising using wavelets. The criterion is based on the minimization of an exactly unbiased estimator of a cost function that gives a measure of similarity between the true unknown model, corresponding with the ideal signal and the candidate model, corresponding with the denoised signal. The cost function considered here is a variant (within a constant) of the Kullback's symmetric divergence, which arguably provides a better assessment of model's similarity than the directed Kullback's divergence. The proposed method called KIC_c-denoising performers very well as compared with classical methods for different type of signals and under different noise levels in terms of minimum MSE of reconstruction. It is also robust by achieving the smallest MSE variance. The good performance of KIC_c for different types of signals make it a an interesting denoising tool in the case where we do not have reliable a priori information about the smoothness of the clean signal. Possible extensions of the proposed method could be directed towards 2D signals (image) denoising, by means of a proper modelling of the problem as a multivariate linear regression.

We did not investigate the performance of the proposed criterion for non Gaussian noise. In the interesting type of non Gaussian noise, i.e., noise with heavy tails, we anticipate that KIC_c-denoising will exhibit a robust behavior as it has a strong penalty function as compared with the MDL. Unlike *Sure* and *Visu*, the computation of KIC_c and MDL do not involve a roughly selected noise level, which is even loose when the noise is tailed. For correlated noise, we anticipate also a robust behavior as compared with the cross-validation principle since this last is based on data splitting which assumes i.i.d distributions. The question, however, whether the proposed method, or more generally model selection based methods are adequate for all the denoising problems, is completely beyond the scope of this paper.

References

[1] D.L. Donoho, I. Johnstone, Ideal spatial adaptation via wavelet shrinkage, Biometrika 81 (1994) 425–455.

[2] D.L. Donoho, I. Johnstone, Adapting unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc. 90 (1995) 1200–1224.

[3] H. Jeffreys, An invariant form of the prior probability in estimation problems, J. Roy. Statist. Soc. A (1946) 453–469.

[4] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.

[5] N. Saito, Simultaneous noise suppression and signal compression using a library of orthogonal bases and the minimum description length criterion, in: E. Foufoula-Georgiou, P. Kumat, (Eds.), Wavelets in Geophysics, 1994, pp. 299–324.

[6] A. Antoniadis, I. Gijbels, G. Grégoire, Model selection using wavelet decomposition and applications, Biometrika 84 (4) (1997) 751–763.
[7] M.H. Hansen, B. Yu, Wavelet thresholding via MDL for natural images, IEEE Trans. Inform. Theory 46 (2000) 1778–1788.
[8] J. Rissanen, MDL denoising, IEEE Trans. Inform. Theory 46 (2000) 2537–2543.
[9] V. Vapnik, The Nature of Statistical Learning Theory, Springer, Berlin, 1995.
[10] V. Cherkassky, X. Shao, Signal estimation and denoising using VC-theory, Neural Networks 14 (2001) 37–52.
[11] H. Linhart, W. Zucchini, Model Selection. Wiley series in Probability and Mathematical Statistics, Wiley, New York, 1986.
[12] A.K. Seghouane, M. Bekara, G. Fleury, A small model selection criterion based on Kullback's symmetric divergence, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong, 2003, pp. 145–148.
[13] S. Kotz, N.L. Johnson, Encyclopedia of Statistical Sciences, vol. 2, Wiley, New York, 1982.
[14] A.K. Seghouane, M. Bekara, A small model selection criterion based on Kullback's symmetric divergence, IEEE Trans. Signal Process. 12 (2004) 3314–3323.
[15] S. Aouada, M. Bekara, A.K. Zoubir, G. Fleury, C.M.S. See, A Gerschgorin–Kullback criterion for source number detection in nonuniform noise and small samples, in: IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), Barcelona, 2004.
[16] M. Bekara, G. Fleury, Bias of the corrected KIC for underfitted regression models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montréal, Canada, 2004, pp. II 517–520.
[17] S.G. Mallat, A Wavelet Tour of Signal Processing, second ed., Academic Press, New York, 1999.
[18] G.P. Nason, Wavelet shrinkage using cross-validation, J. Roy. Statist. Soc. Ser. B 58 (2) (1998) 463–479.
[19] G.P. Nason, B.W. Silverman, The discrete wavelet transform in S, J. Comput. Graphical Statist. 3 (1994) 163–193.