Nonparametric Cepstrum Estimation via Optimal Risk Smoothing

Randy C. S. Lai, Thomas C. M. Lee, *Senior Member, IEEE*, Raymond K. W. Wong, and Fang Yao

***Abstract*—This paper proposes a new cepstrum estimation procedure that is capable of producing smoother and improved cepstrum estimates without the use of any parametric modeling. This procedure consists of two main steps: In the first step, it applies a so-called grid transformation to the empirical cepstral coefficients, while in the second step it nonparametrically smooths the transformed coefficients with local linear regression. The Stein's unbiased risk estimation (SURE) approach is adopted to select both the extent of the grid transformation and the amount of smoothing. It is shown that the use of this SURE selection method for the current problem is asymptotically optimal in a well-defined sense. Lastly, the good practical performance of the new cepstrum estimation procedure is demonstrated via numerical experiments.**

***Index Terms*—Bandwidth selection, grid transformation, local linear regression, Stein's unbiased risk estimation (SURE), thresholding.**

I. Introduction

THE study of cepstrum can be dated as early as [3]. Since then, it has been applied widely in many different areas, including spectral estimation, filter design, image processing and geology, just to name a few; e.g., see [19], [20], [22], and references given therein. As noted by [19] and [25], given the many successful stories of applications of cepstrum, it is almost certain that new and useful applications of cepstrum will emerge. Therefore, it is important to have high-performance procedures for cepstral estimation. The goal of this paper is to propose such a new estimation procedure.

Suppose a finite-sized realization $y_0, \ldots, y_{2n-1}$ of a real-valued, discrete-time, stationary signal $\{y_t\}$ is observed. Denote its power spectral density as Φ, and write

$$\omega_j = \frac{2\pi j}{2n}, \quad \Phi_j = \Phi(\omega_j), \quad j = 0, \ldots, 2n-1.$$

Manuscript received December 29, 2008; accepted September 23, 2009. First published November 06, 2009; current version published February 10, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Patrice Abry. This work was supported in part by a Chinese University of Hong Kong Direct grant, by the Hong Kong Research Grants Council under CERG 401507, by the National Science Foundation under Grant 0707037, and by a Natural Sciences and Engineering Research Council Discovery Grant of Canada.

R. C. S. Lai and R. K. W. Wong are with the Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (e-mail: s0802158@sta.cuhk.edu.hk; s0802162@sta.cuhk.edu.hk).

T. C. M. Lee is with the Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, and also with the Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877 USA (e-mail: tlee@sta.cuhk.edu.hk).

F. Yao is with the Department of Statistics, University of Toronto, Toronto, ON M5S 3G3, Canada (e-mail: fyao@utstat.toronto.edu).

Digital Object Identifier 10.1109/TSP.2009.2036067

It is assumed that $\Phi_j > 0$ for all j. The cepstrum of $\{y_t\}$ is then defined as

$$c_k = \frac{1}{2n} \sum_{j=0}^{2n-1} \ln(\Phi_j) \exp(\mathrm{i}\omega_k j), \quad k = 0, \ldots, 2n-1.$$

It is straightforward to show the symmetry property

$$c_k = c_{2n-k}, \quad k = 1, \ldots, n.$$

Thus for the rest of this paper, we shall focus on the $n+1$ distinct cepstral coefficients $c_0, \ldots, c_n$. It has been observed that, for many practical situations [11], [26], a lot of these cepstral coefficients are either zeros or extremely small in magnitude. In fact, many thresholding-based cepstrum estimation methods were motivated by this observation. In the next section, more will be said about such thresholding methods.

Denote the periodogram of the observed signal as $\hat{\Phi}_j$

$$\hat{\Phi}_j = \frac{1}{2n} \left| \sum_{t=0}^{2n-1} y_t \exp(-\mathrm{i}\omega_j t) \right|^2, \quad j = 0, \ldots, 2n-1.$$

As with Φ_j, $\hat{\Phi}_j$ is also assumed to be positive for all j. A first crude estimate of the cepstrum is given by the *empirical cepstral coefficients* (sometimes also known as *quefrency values*)

$$\hat{c}_k = \frac{1}{2n} \sum_{j=0}^{2n-1} \ln(\hat{\Phi}_j) \exp(\mathrm{i}\omega_k j) + \gamma\delta_k, \quad k = 0, \ldots, n$$

where

$$\delta_k = \begin{cases} 1, & \text{if } k = 0 \\ 0, & \text{otherwise} \end{cases}$$

and $\gamma = 0.577216$ is Euler's constant. It is known that [26], under some regularity conditions and for large n, these empirical cepstral coefficients $\{\hat{c}_k\}_{k=0}^n$ can be well modeled by independent normal random variables with

$$\hat{c}_k \sim N\left(c_k, s_k^2\right) \tag{1}$$

where

$$s_k^2 = \begin{cases} \pi^2/(6n), & \text{if } k = 0, n \\ \pi^2/(12n), & \text{if } k = 1, \ldots, n-1. \end{cases} \tag{2}$$

In the remainder of this paper, we will assume that this distributional property is exact and from which our new cepstrum estimator is built upon. This new estimator is nonparametric in nature and attempts to provide a smoother and better cepstral estimate while avoiding the use of any parametric model.

The rest of this paper is organized as follows. Section II presents the proposed nonparametric cepstrum estimator. In Section III, some theoretical properties of the estimator are established. The empirical properties of the proposed estimator are then evaluated in Section IV via a simulation study. Lastly, concluding remarks are offered in Section V, while technical details are deferred to the Appendix.

II. Proposed Method

Given (1), (2), and the fact that many of the cepstral coefficients are zeros or small, a sensible method for estimating the cepstrum is thresholding (e.g., [26]). That is, the estimate for c_k is set to zero if $\hat{c}_k$ is less than a thresholding value; otherwise, use $\hat{c}_k$ as the estimate. The thresholding value is typically chosen as a multiple of s_k. This thresholding approach is fast and performs reliably for many different types of cepstra. However, if the cepstrum is "smooth" in the sense that $|c_k - c_j|$ is small whenever their "horizontal distance" $|k-j|$ is small, then the thresholding estimation of c_k can be improved upon. It is because one could borrow useful information from the neighboring empirical cepstral coefficients; i.e., $\{\hat{c}_j : |k-j| < d\}$ for some small cutoff distance d. Indeed, our proposed method is motivated by this argument; loosely speaking, it estimates c_k by using a weighted average of all elements in $\{\hat{c}_j : |k-j| < d\}$ for a carefully chosen d.

A. Grid Transformation

Due to the following reason, the proposed method first applies a so-called *grid transformation* to the data before averaging them. For many real-life signals, such as seismic and underwater acoustic channel data [5], [14], a large portion of their cepstral energy is concentrated in the beginning part of their corresponding cepstra. In other words, a typical cepstrum $\{c_k\}$ has large values and changes more rapidly at its left end, while its right tail is relatively long and flat. This suggests that a smaller d should be used when k is small and a larger d should be used for large values of k.

The same effect can be more conveniently achieved by applying a grid transformation and use the same d for all values of k; e.g., see [12, Sec. 2.3.3]. For simplicity, call the "horizontal distance" of $\hat{c}_k$ from the origin the *x-coordinate*. Therefore the x-coordinate of $\hat{c}_k$ is k, and the whole empirical cepstrum can be plotted by tracing the points $\{(k, \hat{c}_k)\}_{k=0}^{n}$ in the xy plane. Now, the grid transformation is to rescale these x-coordinates so that the horizontal distance between $\hat{c}_k$ and $\hat{c}_{k+1}$ becomes larger for small values of k and smaller for large k.

Such a grid transformation can be accomplished by applying a function $g_r : [0, n] \to [0, n]$ to the x-coordinates of c_k. The subscript r is used to denote a tuning parameter that controls the extent of the transformation; more will be said about this below. This function g_r should be strictly increasing and concave, and it has the identify function as its special case for a particular value of r. We have investigated the use of different g_r's that satisfy these conditions, including $g_r(k) = n(k/n)^r$ and $g_r(k) = n\log(1+rk)/\log(1+rn)$. Our extensive numerical experience suggested that the choice of g_r is not crucial, although the following choice occasionally provided better results:

$$k' = g_r(k) = n\left\{\frac{1-\exp(-rk)}{1-\exp(-rn)}\right\}.$$

For this reason, we shall use this choice of g_r in the rest of this paper; see Fig. 1 for an example illustrating the effect of r. We will discuss the choice of r in Section II-C. With this grid transformation, one could imagine that the set of original empirical cepstral coefficients has been transformed from $\{(k, \hat{c}_k)\}_{k=0}^{n}$ to $\{(k', \hat{c}_k)\}_{k=0}^{n}$. The next step is to smooth $\{(k', \hat{c}_k)\}_{k=0}^{n}$ nonparametrically using local linear regression. The rationale behind this local smoothing is that if the cepstrum is locally smooth, estimation of c_k can be improved by borrowing information from neighboring $\hat{c}_k$'s.

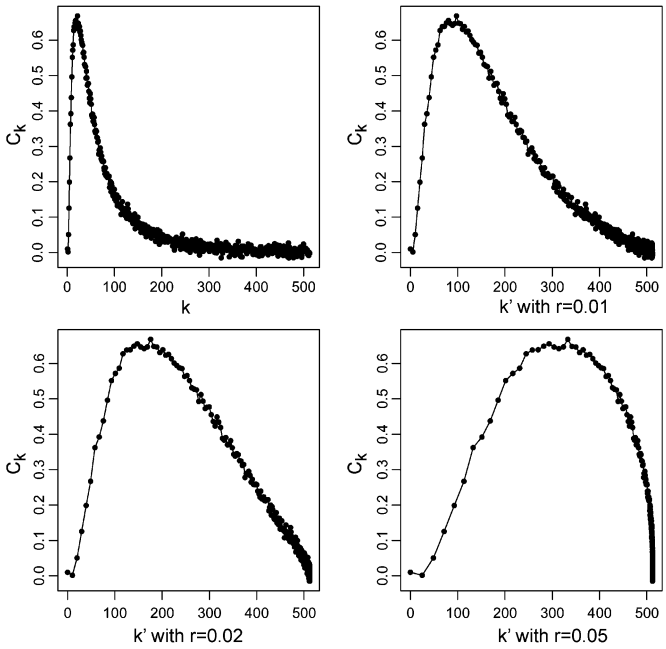


Fig. 1. An artificial empirical cepstrum $\{(k, \hat{c}_k)\}$'s (top left panel) and different grid-transformed cepstra $\{(k', \hat{c}_k)\}$'s with different values of r. One can see that for the artificial cepstrum displayed in the top left panel, a small d is required to avoid oversmoothing at the left end, while a large d is required to stabilize the noisy right tail. However, for a suitably grid-transformed cepstrum, the same d can be applied to all regions.

B. Smoothing Using Local Linear Regression

Denote the local linear regression estimate of c_k as $\tilde{c}_k$. For each k, $\tilde{c}_k$ is calculated by performing a weighted least squares regression using the $\{\hat{c}_j\}_{j=0}^{n}$ as the response and $\{(k'-j')\}_{j=0}^{n}$ as the predictor. The weights are given by $\{K_h(k'-j')\}_{j=0}^{n}$, where $K(\cdot)$ is known as the kernel function, h is the bandwidth that controls the amount of smoothing, and $K_h(\cdot) = (1/h)K(\cdot/h)$. The bandwidth h plays the same role as the cutoff distance d mentioned previously. It is known that [10, Sec. 3.2] as long as $K(\cdot)$ is unimodal and symmetric about 0, its exact form is relatively unimportant. In all our numerical work to be reported below, $K(\cdot)$ is taken as the standard normal density.

The estimate $\tilde{c}_k$ is defined as the intercept $\hat{a}_k$ of the best fitting regression line $\hat{a}_k + \hat{b}_k k'$ that minimizes the following weighted residual sum of squares

$$\sum_{j=0}^{n} \{\hat{c}_j - a_k - b_k(k' - j')\}^2 K_h(k' - j').$$

The minimizers of the above are shown to be

$$(\hat{a}_k, \hat{b}_k)^T = \left(\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k\right)^{-1} \mathbf{X}_k^T \mathbf{W}_k \hat{\mathbf{c}} \tag{3}$$

where $\hat{\mathbf{c}} = (\hat{c}_0, \ldots, \hat{c}_n)^T$, $\mathbf{X}_k$ is a $(n+1)$-by-2 matrix with the jth row as $(1, k' - j')$, and $\mathbf{W}_k$ is a diagonal matrix with diagonal elements $K_h(k' - 0'), \ldots, K_h(k' - n')$. For further details on local linear regression, see [10] and [27] for examples.

Since both $\mathbf{X}_k$ and $\mathbf{W}_k$ are independent of $\hat{c}_k$'s, from (3) we can see that $\hat{a}_k$, or equivalently $\tilde{c}_k$, is a linear combination of the $\hat{c}_k$'s. Thus, we can write

$$\tilde{c}_k = \hat{a}_k = \sum_{j=0}^{n} S_{k,j} \hat{c}_j \tag{4}$$

for some $S_{k,j}$'s that are independent of the $\hat{c}_k$'s; these $S_{k,j}$'s will be used in the next subsection. In other words, if $\mathbf{S}$ is the matrix with $S_{k,j}$ as its (k,j)th element, then (4) can be expressed as

$$\tilde{\mathbf{c}} = \mathbf{S}\hat{\mathbf{c}}$$

with $\tilde{\mathbf{c}} = (\tilde{c}_0, \ldots, \tilde{c}_n)^T$. The matrix $\mathbf{S}$ is sometimes known as the smoothing matrix.

We close this subsection by noting that the above estimate $\tilde{c}_k$ for c_k is a function of the transformation parameter r and the bandwidth h, but for clarity this dependence has been suppressed in the notation of $\tilde{c}_k$. To use $\{\tilde{c}_k\}_{k=0}^n$ as a cepstrum estimator, one needs to choose (h, r). We have developed such an automatic selection method, to be described next.

C. Stein's Unbiased Risk Estimation

A reasonable choice for (h, r) is the pair that jointly minimizes the following risk function:

$$R(h, r) = \mathrm{E}\left\{\frac{1}{n+1}\sum_{k=0}^{n}(\tilde{c}_k - c_k)^2\right\}. \tag{5}$$

Of course, in practice, $R(h, r)$ is an unknown quantity, so a direct minimization is not possible. A common approach to overcome this issue is to construct an unbiased estimator for $R(h, r)$ and choose (h, r) as the minimizer of the resulting estimator. This approach is commonly known as Stein's unbiased risk estimation (SURE) [24] (see also [23] for a more elaborated discussion). It has been successfully used for tackling different problems, such as wavelet thresholding [1], [8], spectral density estimation [15], [28], and image denoising [2], [4], [21]. For generalizations of SURE, see [9] and [13] for examples.

For the current cepstrum smoothing problem, we have derived an approximate unbiased estimator for $R(h, r)$. This estimator is exactly unbiased if (1) and (2) were true. We propose to choose (h, r) as its joint minimizer. The expression of this estimator is given below, and the justification for its unbiasedness under (1) and (2) is provided in Appendix A.

Theorem 1: Under (1) and (2), the risk estimator $\hat{R}(h, r)$ defined in (6) is an unbiased estimator of $R(h, r)$. That is

$$E\left\{\hat{R}(h, r)\right\} = R(h, r)$$

where

$$\hat{R}(h, r) = \frac{1}{n+1}\left\{\sum_{k=0}^{n}(\tilde{c}_k - \hat{c}_k)^2 - (1 - S_{0,0} - S_{1,1})\frac{\pi^2}{3n} - \sum_{k=1}^{n-1}(1 - 2S_{k,k})\frac{\pi^2}{12n}\right\}. \tag{6}$$

To sum up, our proposed estimator is defined by (4), with (h, r) chosen as the minimizer of (6). Below, we refer to this estimator as *SURESmooth*.

D. Minimization of $\hat{R}(h, r)$

A straightforward but also time-consuming method to minimize $\hat{R}(h, r)$ with respect to (h, r) is to conduct a two-dimensional grid search. For $n = 512$, if the search was performed on a 20×20 grid of (h, r), our implementation requires around 10 s to finish with a Core2Duo 2.4 GHz processor. This may not be fast enough for many real problems. However, we have observed that, for many different data sets, the surfaces of $\hat{R}(h, r)$ are smooth when plotted against h and r. This suggests that many simple strategies should work well for speeding up the minimization of $\hat{R}(h, r)$. We have used the following.

The idea behind our strategy is to decompose the two-dimensional search into a sequence of one-dimensional searches. First, we fix a value for h at, say, h_0, and find the corresponding value of r that minimizes $\hat{R}(h_0, r)$. Denote this value of r as r_1. Then, we set r as r_1 and find the value of h so that $\hat{R}(h, r_1)$ is minimized. Denote this value of h as h_1, and next we find the value of r that minimizes $\hat{R}(h_1, r)$. We keep iterating this process until the value of $\hat{R}(h, r)$ cannot be made smaller. When comparing to more classical methods such as Newton–Raphson, one attractive property of this strategy is that no calculation is needed for the gradient or higher derivatives of $\hat{R}(h, r)$. On average, this procedure takes about 2 s to finish with the same machine mentioned above. When comparing to many fast cepstrum estimation procedures such as the thresholding method of [26], our approach is still computationally more expensive; e.g., our implementation of [26] on average takes about 0.005 s to finish, making our method about 400 times slower. However, the potential improvement in estimation quality do make our approach a viable alternative.

III. Asymptotic Optimality of *SURESmooth*

In this section, we study the theoretical properties of *SURESmooth*. To be more specific, we shall show that the use of the unbiased risk estimator $\hat{R}(h, r)$ (6) for choosing (h, r) is asymptotically optimal in a well-defined sense, as stated in (8).

For technical simplicity, we shall assume that the risk estimator $\hat{R}(h, r)$ is minimized over a discrete index set $\mathcal{A}_n$. In other words, its joint minimizer is restricted to be an element of $\mathcal{A}_n$, where $\mathcal{A}_n$ can be seen as a two-dimensional gridded value

of (h,r). The grid density of $\mathcal{A}_n$ could always be made sufficiently dense enough so that in practice there is virtually no difference if $\hat{R}(h,r)$ is minimized over $\mathcal{A}_n$ or $\mathcal{R}^2$. We shall denote the order of the cardinality of $\mathcal{A}_n$ as n^δ; i.e., $|\mathcal{A}_n| = O(n^\delta)$ for some $\delta > 0$.

Define the loss function for the *SURESmooth* estimator as

$$L(h,r) = \frac{1}{n+1}\sum_{k=0}^{n}(\tilde{c}_k - c_k)^2 = \frac{1}{n+1}\|\tilde{\mathbf{c}} - \mathbf{c}\|^2 \quad (7)$$

where $\mathbf{c} = (c_0, \ldots, c_n)^T$. Let $(\hat{h}, \hat{r}) \in \mathcal{A}_n$ be the minimizer of (6). Then, our proposed selection procedure is asymptotically optimal in the following sense:

$$\frac{L(\hat{h},\hat{r})}{\inf_{(h,r)\in\mathcal{A}_n} L(h,r)} \xrightarrow{P} 1. \quad (8)$$

Similar definitions for asymptotic optimality have also been studied by previous authors in different contexts, for both parametric and nonparametric model selection problems (e.g., [7], [16], [17]).

Recall that $\mathbf{S}$ is the $n \times n$ smoothing matrix; i.e., $\tilde{\mathbf{c}} = \mathbf{S}\hat{\mathbf{c}}$. Denote the maximum singular value of $\mathbf{S}$ by $\lambda_{\mathbf{S}}$. The assumptions required for establishing the above asymptotic optimality of $\hat{R}(h,r)$ are:

(A1) $\limsup_{n\to\infty} \sup_{(h,r)\in\mathcal{A}_n} \lambda_{\mathbf{S}} < \infty$;

(A2) $\sum_{(h,r)\in\mathcal{A}_n} \{n^2 R(h,r)\}^{-1} \longrightarrow 0$.

Assumption (A1) is natural, and in fact if $\lambda_{\mathbf{S}} > 1$, then $\tilde{\mathbf{c}}$ is inadmissible and dominated by some other linear estimators [6]. To understand (A2), one first notes that the optimal risk $R(h,r)$ is typically of order $n^{-\delta'}$ for some $\delta' > 0$ as $n \to \infty$. If the cardinality of $\mathcal{A}_n$ is of polynomial order n^δ, one can see that the upper bound on its magnitude $\delta < 2 - \delta'$ will usually allow a sufficient grid search in practice.

The following theorem summarizes the aforementioned desirable theoretical property of the proposed selection method. The proof is deferred to Appendix B.

Theorem 2: Under (1) and (2), the risk estimator $\hat{R}(h,r)$ is asymptotically optimal under assumptions (A1)–(A2). That is, (8) holds for $(\hat{h},\hat{r}) = argin f_{(h,r)\in\mathcal{A}_n}\hat{R}(h,r)$.

IV. Simulation Results

A simulation study has been conducted to evaluate the empirical performance of *SURESmooth*. The following four models were used for generating the testing signals $\{y_t\}$:

- *Model 1*: a broadband MA with a medium dynamic range log-spectrum

$$y_t = e_t + 0.4574e_{t-1} + 0.2157e_{t-2} + 0.35951e_{t-3} + 0.1383e_{t-4}.$$

- *Model 2*: A narrowband ARMA with a large dynamic range log-spectrum

$$y_t = 1.55y_{t-1} - 0.95y_{t-2} + e_t + 0.75e_{t-1} + 0.35e_{t-2}.$$

- *Model 3*: A broadband AR with a small dynamic range log-spectrum

$$y_t = 1.5y_{t-1} - 0.7y_{t-2} + 0.1y_{y-3} + e_t.$$

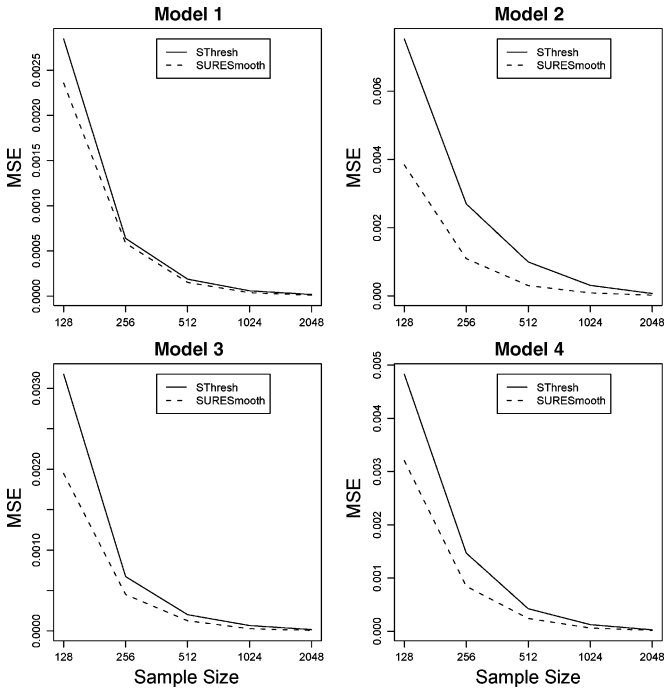


Fig. 2. Simulation results: MSE averages in the cepstrum domain for Model 1 (top left panel), Model 2 (top right panel), Model 3 (bottom left panel), and Model 4 (bottom right panel).

- *Model 4*: A broadband MA with a medium dynamic range log-spectrum

$$y_t = e_t - 0.3e_{t-1} - 0.6e_{t-2} - 0.3e_{t-3} + 0.6e_{t-4}.$$

In the above, the e_t's are iid $N(0,1)$ white noise. Models 1 and 2 have been used by [26] in the context of cepstrum thresholding, while the remaining two models have been used by various researchers for spectrum smoothing (e.g., [15] and [18]). We considered five different sample sizes: $2n = 2^l$ for $l = 7$ to 11. For each combination of model and sample size, 500 realizations of $\{y_t\}$ were generated. Then, *SURESmooth* was applied to each generated realization to obtain the estimate $\tilde{c}_k$ for c_k. The following mean squared error (MSE) was computed as a measure for quality of fit:

$$\text{MSE} = \frac{1}{n+1}\sum_{k=0}^{n}(\tilde{c}_k - c_k)^2.$$

For comparison purposes, the *SThresh* method of [11] and [26] was also applied to all generated $\{y_t\}$ to estimate c_k, and the corresponding MSEs were also computed.

The averages of the computed MSEs for different combinations of model, sample size, and estimation method are summarized in Fig. 2. From these plots, one could see that the proposed method *SURESmooth* always gave smaller MSE averages than *SThresh*. We have also applied paired t-tests to these MSE values, and the results show that the MSE average differences are statistically significant.

Since very often a major goal of cepstral analysis is to perform spectrum or log-spectrum estimation, for each set of esti-

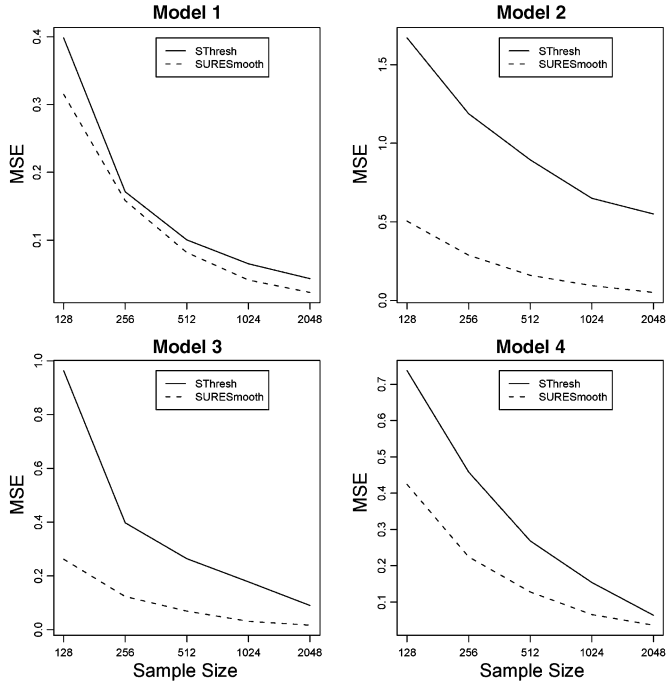


Fig. 3. Simulation results: MSE averages in the log-spectrum domain for Model 1 (top left panel), Model 2 (top right panel), Model 3 (bottom left panel), and Model 4 (bottom right panel).

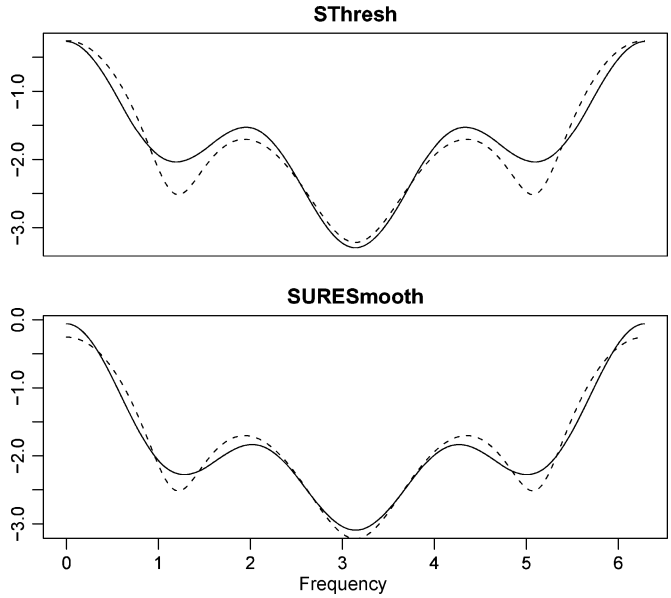


Fig. 4. Estimated log-spectra of Model 1 obtained from *SThresh* and *SURESmooth* with $2n = 2048$. In both panels, the solid line represents the 250th smallest MSE estimated log-spectrum, while the dotted line is the true log-spectrum.

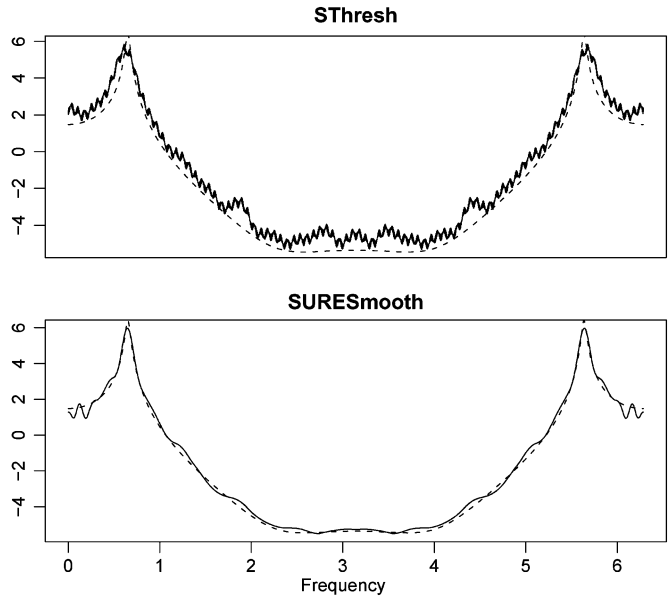


Fig. 5. Similar to Fig. 4, but for Model 2.

mated cepstrum $\tilde{c}_k$, we also calculated the corresponding estimated log-spectrum

$$\ln(\tilde{\Phi}_j) = \sum_{k=0}^{2n-1} \tilde{c}_k \exp(-\mathrm{i}k\omega_j), \quad j = 0, \ldots, n$$

and its MSE, defined as

$$\frac{1}{n+1} \sum_{j=0}^{n} \left\{ \ln(\tilde{\Phi}_j) - \ln(\Phi_j) \right\}^2.$$

The averages of these MSEs are displayed in Fig. 3 in a similar fashion as Fig. 2. Once again, *SURESmooth* seems to be a preferred method.

To visually evaluate the quality of the fitted log-spectra, we ranked the 500 *SURESmooth* MSEs that correspond to the combination of Model 1 and $2n = 2048$. The estimated log-spectrum that has the 250th smallest MSE is shown in Fig. 4. Similar plots were obtained for Model 2 to Model 4, and for the *SThresh* method; see Figs. 4 to 7. These plots seem to suggest that those *SURESmooth* estimates tend to be superior to those from *SThresh*.

V. Concluding Remarks

In this paper, a new and automatic method for cepstrum estimation, *SURESmooth*, is presented. This method is nonparametric and capable of producing smoother and better cepstrum estimates without imposing any parametric model. The two main ingredients of *SURESmooth* are grid transformation and local linear smoothing. The tuning parameters of *SURESmooth* are chosen automatically by Stein's unbiased risk estimation approach. It is theoretically shown that this parameter choice is asymptotically optimal. In addition, simulation results suggest that *SURESmooth* can be a better alternative for estimating both cepstrum and log-spectrum.

Appendix A
Proof of Theorem 1: Unbiasedness of $\hat{R}(h, r)$

This appendix outlines the derivation of the risk estimator (6). We iterate again that this estimator is developed under (1) and (2).

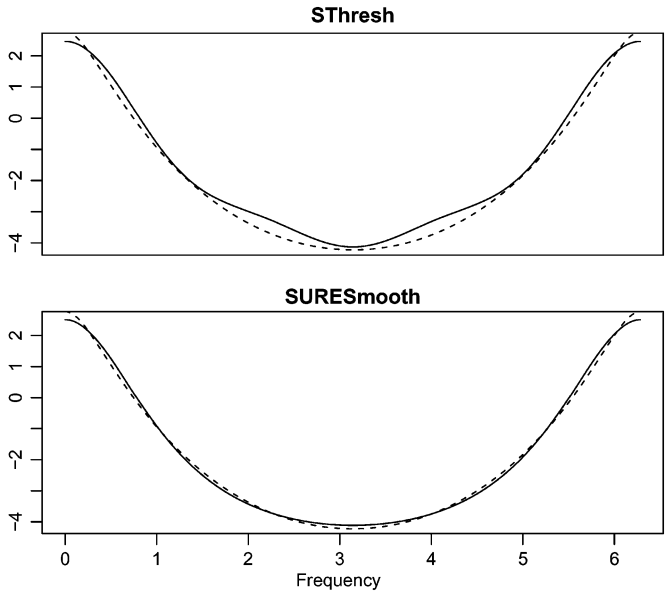


Fig. 6. Similar to Fig. 4, but for Model 3.

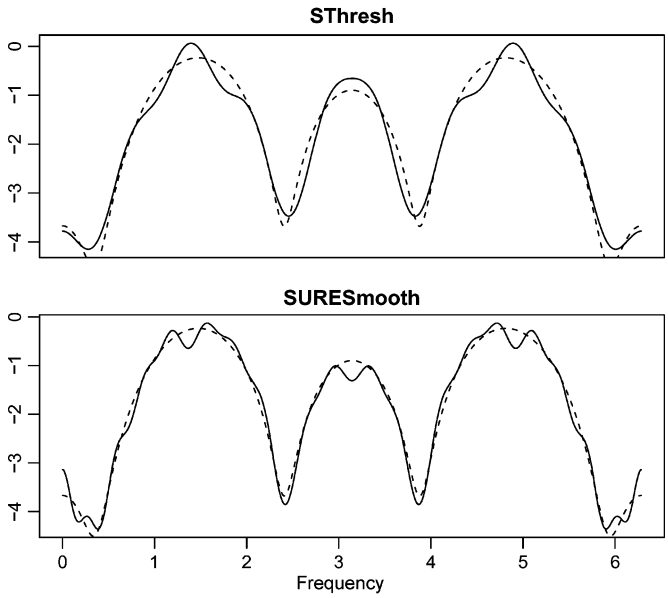


Fig. 7. Similar to Fig. 4, but for Model 4.

First, we calculate

$$\mathrm{E}\left\{(\tilde{c}_k-\hat{c}_k)^2\right\}=\mathrm{E}\left\{(\tilde{c}_k-c_k)^2\right\}+\mathrm{E}\left\{(\hat{c}_k-c_k)^2\right\} -2\mathrm{E}\left\{(\tilde{c}_k-c_k)(\hat{c}_k-c_k)\right\}. \quad (9)$$

Note that the second term on the right-hand side is $\mathrm{Var}(\hat{c}_k)$. Using $\tilde{c}_k=\sum_{j=0}^{n}S_{k,j}\hat{c}_j$ from (4) and $\mathrm{E}(\hat{c}_k)=c_k$, the last term can be calculated as

$$\begin{aligned}\mathrm{E}\left\{(\tilde{c}_k-c_k)(\hat{c}_k-c_k)\right\}&=\mathrm{E}\left\{\tilde{c}_k(\hat{c}_k-c_k)\right\}-\mathrm{E}\left\{c_k(\hat{c}_k-c_k)\right\}\\&=\mathrm{E}\left\{\sum_{j=0}^{n}S_{k,j}\hat{c}_j(\hat{c}_k-c_k)\right\}-0.\end{aligned}$$

As $\mathrm{E}\{\hat{c}_j(\hat{c}_k-c_k)\}=0$ whenever $j\neq k$, the previous calculation becomes

$$\begin{aligned}\mathrm{E}\left\{(\tilde{c}_k-c_k)(\hat{c}_k-c_k)\right\}&=\mathrm{E}\left\{S_{k,k}\hat{c}_k(\hat{c}_k-c_k)\right\}\\&=S_{k,k}\left[\mathrm{E}\left(\hat{c}_k^2\right)-\left\{\mathrm{E}(\hat{c}_k)\right\}^2\right]\\&=S_{k,k}\mathrm{Var}(\hat{c}_k).\end{aligned}$$

Now, summing (9) over k and dividing by $n+1$, we have

$$\begin{aligned}&\frac{1}{n+1}\sum_{k=0}^{n}\mathrm{E}\left\{(\tilde{c}_k-\hat{c}_k)^2\right\}\\&\qquad=R(h,r)+\frac{1}{n+1}\sum_{k=0}^{n}\left\{\mathrm{Var}(\hat{c}_k)-2S_{k,k}\mathrm{Var}(\hat{c}_k)\right\}.\end{aligned}$$

Replacing the expectation operation with summation, we establish that

$$\hat{R}(h,r)=\frac{1}{n+1}\left\{\sum_{k=0}^{n}(\tilde{c}_k-\hat{c}_k)^2-\sum_{k=0}^{n}(1-2S_{k,k})\mathrm{Var}(\hat{c}_k)\right\}$$

is an unbiased estimator for $R(h,r)$ under (1) and (2). The estimator (6) can then be straightforwardly obtained by replacing $\mathrm{Var}(\hat{c}_k)$ with the corresponding values given in (1) and (2).

Appendix B
Proof of Theorem 2: Asymptotic Optimality of $\hat{R}(h,r)$

This appendix presents the proof for Theorem 2. From the derivation in Appendix A, $s_k^2=\mathrm{Var}(\hat{c}_k)$ given in (2) does not depend on (h,r), hence minimizing $\hat{R}(h,r)$ (6) is equivalent to minimizing

$$\tilde{R}(h,r)=\frac{1}{(n+1)}\left\{\sum_{k=0}^{n}(\tilde{c}_k-\hat{c}_k)^2+2\sum_{k}S_{k,k}s_k^2\right\}.$$

For convenience, we shall deal with $\tilde{R}(h,r)$ in the remainder of this section and abbreviate $\sum_{k=0}^{n}$ as $\sum_k$. Let $\epsilon_k=\hat{c}_k-c_k$, and thus $\epsilon_k\sim N(0,s_k^2)$. Write $\boldsymbol{\epsilon}=(\epsilon_0,\ldots,\epsilon_n)^T$ and $\mathbf{A}=\mathbf{I}-\mathbf{S}$, where $\mathbf{I}$ is the $n\times n$ identity matrix. We have

$$\begin{aligned}(n+1)\tilde{R}(h,r)&=\|\hat{\mathbf{c}}-\mathbf{c}+\mathbf{c}-\tilde{\mathbf{c}}\|^2+2\sum_{k}S_{k,k}s_k^2\\&=\|\boldsymbol{\epsilon}\|^2+(n+1)L(h,r)+2\langle\boldsymbol{\epsilon},\mathbf{c}-\tilde{\mathbf{c}}\rangle\\&\quad+2\sum_{k}S_{k,k}s_k^2\\&=\|\boldsymbol{\epsilon}\|^2+(n+1)L(h,r)+2\langle\boldsymbol{\epsilon},\mathbf{A}\mathbf{c}\rangle\\&\quad+2\left(\sum_{k}S_{k,k}s_k^2-\langle\boldsymbol{\epsilon},\mathbf{S}\boldsymbol{\epsilon}\rangle\right). \quad (10)\end{aligned}$$

Since $\|\boldsymbol{\epsilon}\|^2$ does not depend on (h,r), in order to prove (8), it is sufficient to show that

$$\sup_{\mathcal{A}_n}\left|\langle\boldsymbol{\epsilon},\mathbf{A}\mathbf{c}\rangle\right|/\left\{nR(h,r)\right\}\xrightarrow{P}0, \quad (11)$$

$$\sup_{\mathcal{A}_n}\left|\sum_{k}S_{k,k}s_k^2-\langle\boldsymbol{\epsilon},\mathbf{S}\boldsymbol{\epsilon}\rangle\right|/\left\{nR(h,r)\right\}\xrightarrow{P}0 \quad (12)$$

and

$$\sup_{\mathcal{A}_n} |L(h,r)/R(h,r) - 1| \xrightarrow{P} 0. \tag{13}$$

To show (11), we apply Chebyshev's inequality: For any $\tau_1 > 0$, noting $s_k^2 \sim n^{-1}$, one has

$$\begin{aligned} P\left\{\sup_{\mathcal{A}_n} \frac{|\langle \boldsymbol{\epsilon}, \mathbf{Ac}\rangle|}{nR(h,r)} > \tau_1\right\} &\le \frac{1}{\tau_1^2} \sum_{\mathcal{A}_n} \frac{E\left(\langle \boldsymbol{\epsilon}, \mathbf{Ac}\rangle^2\right)}{n^2 R(h,r)^2} \\ &\le \frac{C_1}{\tau_1^2} \sum_{\mathcal{A}_n} \frac{\|\mathbf{Ac}\|^2}{n^3 R^2(h,r)} \end{aligned} \tag{14}$$

for some constant $C_1 > 0$. Since $nR(h,r) = \|\mathbf{Ac}\|^2 + E(\|\mathbf{S}\boldsymbol{\epsilon}\|^2) \ge \|\mathbf{Ac}\|^2$ and (A2), the right-hand side of (14) is bounded by $C_1 \tau_1^{-2} \sum_{\mathcal{A}_n} \{n^2 R(h,r)\}^{-1} \to 0$. Then, (11) is proved.

Equation (12) can be shown by observing

$$E\langle \boldsymbol{\epsilon}, \mathbf{S}\boldsymbol{\epsilon}\rangle = E\left(\sum_k S_{k,k}\epsilon_k^2\right) = \sum_k S_{k,k} s_k^2$$

and

$$nR(h,r) \ge E\left(\|\mathbf{A}\boldsymbol{\epsilon}\|^2\right) = E\left(\sum_k \sum_j S_{k,j}^2 \epsilon_k^2\right) \ge \frac{C_2}{n} \mathrm{tr}(\mathbf{S}^T\mathbf{S})$$

for some $C_2 > 0$. It is easy to check that, for normal random vector $\boldsymbol{\epsilon} = \hat{\mathbf{c}} - \mathbf{c}$ distributed as in (1), $\mathrm{Var}(\langle \boldsymbol{\epsilon}, \mathbf{S}\boldsymbol{\epsilon}\rangle) \sim n^{-2} tr(\mathbf{S}^T\mathbf{S})$. Then, (12) follows by (A2), for any $\tau_2 > 0$ and some $C_3 > 0$

$$\begin{aligned} &P\left\{\sup_{\mathcal{A}_n} \frac{\left|\sum_k S_{k,k} s_k^2 - \langle \boldsymbol{\epsilon}, \mathbf{S}^T\boldsymbol{\epsilon}\rangle\right|}{nR(h,r)} > \tau_2\right\} \\ &\le \frac{1}{\tau_2^2} \sum_{\mathcal{A}_n} \frac{E\left\{\left(\langle \boldsymbol{\epsilon}, \mathbf{S}\boldsymbol{\epsilon}\rangle - E\langle \boldsymbol{\epsilon}, \mathbf{S}\boldsymbol{\epsilon}\rangle\right)^2\right\}}{n^2 R^2(h,r)} \\ &\le \frac{C_3}{\tau_2^2} \sum_{\mathcal{A}_n} \frac{\mathrm{tr}(\mathbf{S}^T\mathbf{S})}{n^4 R^2(h,r)} \\ &\le \frac{C_3}{\tau_2^2} \sum_{\mathcal{A}_n} \frac{1}{n^2 R(h,r)} \longrightarrow 0. \end{aligned}$$

To show (13), one notes

$$L(h,r) - R(h,r) = \frac{1}{n}\left\{\|\mathbf{S}\boldsymbol{\epsilon}\|^2 - E\left(\|\mathbf{S}\boldsymbol{\epsilon}\|^2\right) - 2\langle \mathbf{Ac}, \mathbf{S}\boldsymbol{\epsilon}\rangle\right\}.$$

Then, it is sufficient to show

$$\sup_{\mathcal{A}_n} |\langle \mathbf{Ac}, \mathbf{S}\boldsymbol{\epsilon}\rangle| / \{nR(h,r)\} \xrightarrow{P} 0 \tag{15}$$

and

$$\sup_{\mathcal{A}_n} \left|\|\mathbf{S}\boldsymbol{\epsilon}\|^2 - E\left(\|\mathbf{S}\boldsymbol{\epsilon}\|^2\right)\right| / \{nR(h,r)\} \xrightarrow{P} 0 \tag{16}$$

which is similar to the proofs of (11) and (12). Observing

$$\langle \mathbf{Ac}, \mathbf{S}\boldsymbol{\epsilon}\rangle = \langle \boldsymbol{\epsilon}, \mathbf{S}^T\mathbf{Ac}\rangle \quad \text{and} \quad \|\mathbf{S}^T\mathbf{Ac}\|^2 \le \lambda_{\mathbf{S}} \|\mathbf{Ac}\|^2$$

that leads to (15), while

$$\|\mathbf{S}\boldsymbol{\epsilon}\|^2 = \langle \boldsymbol{\epsilon}, \mathbf{S}^T\mathbf{S}\boldsymbol{\epsilon}\rangle \quad \text{and} \quad \mathrm{tr}(\mathbf{S}^T\mathbf{S}\mathbf{S}^T\mathbf{S}) \le \lambda_{\mathbf{S}}^2 \mathrm{tr}(\mathbf{S}^T\mathbf{S})$$

completes the proof of (16).

ACKNOWLEDGMENT

The authors are most grateful to the reviewers for many useful comments and suggestions.

REFERENCES

[1] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Stein's principle," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1814–1830, Nov. 2005.

[2] T. Blu and F. Luisier, "The SURE-LET approach to image denoising," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2778–2786, Nov. 2007.

[3] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proc. Symp. Time Series Anal.*, 1963, pp. 209–243.

[4] C. Chaux, L. Duval, A. Benazza-Benyahia, and J.-C. Pesquet, "A nonlinear Stein-based estimator for multichannel image denoising," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pt. 2, pp. 3855–3870, Aug. 2008.

[5] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proc. IEEE*, vol. 65, no. 10, pp. 1428–1443, Oct. 1977.

[6] A. Cohen, "All admissible linear estimators of the mean vector," *Ann. Math. Stat.*, vol. 37, pp. 458–463, 1966.

[7] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, pp. 377–403, 1979.

[8] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, 1995.

[9] Y. C. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 471–481, Feb. 2009.

[10] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*. London, U.K.: Chapman & Hall, 1996.

[11] E. Gudmundson, N. Sandgren, and P. Stoica, "Automatic smoothing of periodograms," in *Proc. 31st ICASSP*, 2006, vol. 3, pp. III: 504–III: 507.

[12] J. D. Hart, *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer, 1997.

[13] H.-C. Huang and T. C. M. Lee, "Data adaptive median filters for signal and image denoising using a generalized SURE criterion," *IEEE Signal Process. Lett.*, vol. 13, no. 9, pp. 561–564, Sep. 2006.

[14] L. LeBlanc, "Narrow-band sampled-data techniques for detection via the underwater acoustic communication channel," *IEEE Trans. Commun. Technol.*, vol. CT-17, no. 4, pp. 481–488, Aug. 1969.

[15] T. C. M. Lee, "A simple span selector for periodogram smoothing," *Biometrika*, vol. 84, pp. 965–969, 1997.

[16] K. C. Li, "Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing," *Ann. Stat.*, vol. 14, pp. 1101–1112, 1986.

[17] K. C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *Ann.f Stat.*, vol. 15, pp. 958–975, 1987.

[18] H. C. Ombao, J. A. Raz, R. L. Strawderman, and R. von Sachs, "A simple generalised cross-validation method of span selection for periodogram smoothing," *Biometrika*, vol. 88, pp. 1186–1192, 2001.

[19] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 95–106, Sep. 2004.

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[21] S. Ramani, T. Blu, and M. Unser, "Monte-Carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1540–1554, Sep. 2008.

[22] V. Solo, "Modeling of two-dimensional random fields by parametric cepstrum," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 6, pp. 743–750, Nov. 1986.
[23] V. Solo, "A sure-fired way to choose smoothing parameters in ill-conditioned inverse problems," in *Proc. IEEE 1996 Int. Conf. Image Process.*, 1996, vol. 3, pp. 89–92.
[24] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, pp. 1135–1151, 1981.
[25] P. Stoica and N. Sandgren, "Smoothed nonparametric spectral estimation via cepstrum thresholding," *IEEE Signal Process. Mag.*, vol. 23, no. 6, pp. 34–45, Nov. 2006.
[26] P. Stoica and N. Sandgren, "Total-variance reduction via thresholding: Application to cepstral analysis," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 66–72, Jan. 2007.
[27] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman & Hall, 1995.
[28] F. Yao and T. C. M. Lee, "Spectral density estimation using sharpened periodograms," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4711–4716, Sep. 2007.

Randy C. S. Lai received the B.Sc. degree in statistics with first-class honors from the Chinese University of Hong Kong (CUHK), Hong Kong, China, in 2008. He is currently pursuing the M.Phil. degree in statistics at CUHK.

His current research interests include statistical learning and digital signal processing.

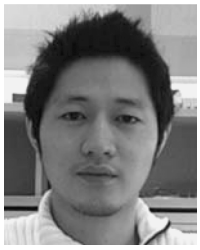
Thomas C. M. Lee (M'97–SM'05) received the B.App.Sci. degree in mathematics and the B.Sc. (Hons.) degree in mathematics with the University Medal from the University of Technology, Sydney, Australia, in 1992 and 1993, respectively, and the Ph.D. degree jointly from Macquarie University and CSIRO Mathematical and Information Sciences, Sydney, Australia, in 1997.

Previously, he has held faculty positions at Colorado State University, Boulder, and the Chinese University of Hong Kong, Hong Kong, China. Currently, he is a Professor of Statistics with the University of California at Davis. His research interests include computational statistics, wavelet techniques, and digital signal and image processing.

Prof. Lee is a Fellow of the American Statistical Association.

Raymond K. W. Wong received the B.Sc. degree in statistics with first-class honors from the Chinese University of Hong Kong (CUHK), Hong Kong, China, in 2008. He is currently pursuing the M.Phil. degree in statistics at CUHK.

His current research interests include smoothing methods, machine learning, and signal processing.



Fang Yao received the Ph.D. degree in statistics from the University of California at Davis in 2003.

Currently, he is an Associate Professor with the Department of Statistics, University of Toronto, Toronto, ON, Canada. His research interests are in functional data analysis, longitudinal studies, nonparametric regression, and applying complex statistical methods to science in general.