Ideal Denoising in an orthonormal basis chosen from a library of bases

David L. Donoho & Iain M. Johnstone
Department of Statistics
Stanford University

Abstract

Suppose we have observations $y_i = s_i + z_i$, $i = 1, ..., n$, where (s_i) is signal and (z_i) is i.i.d. Gaussian white noise. Suppose we have available a library $\mathcal{L}$ of orthogonal bases, such as the Wavelet Packet bases or the Cosine Packet bases of Coifman and Meyer. We wish to select, adaptively based on the noisy data (y_i), a basis in which best to recover the signal ("de-noising"). Let M_n be the total number of distinct vectors occcuring among all bases in the library and let $t_n = \sqrt{2\log(M_n)}$. (For wavelet packets, $M_n = n \log_2(n)$.)

Let $y[\mathcal{B}]$ denote the original data y transformed into the Basis $\mathcal{B}$. Choose $\lambda > 8$ and set $\Lambda_n = (\lambda \cdot (1 + t_n))^2$. Define the entropy functional

$$\mathcal{E}_\lambda(y, \mathcal{B}) = \sum_i \min(y_i^2[\mathcal{B}], \Lambda_n^2).$$

Let $\hat{\mathcal{B}}$ be the best orthogonal basis according to this entropy:

$$\hat{\mathcal{B}} = \arg\min_{\mathcal{B} \in \mathcal{L}} \mathcal{E}_\lambda(y, \mathcal{B}).$$

Define the hard-threshold nonlinearity $\eta_t(y) = y 1_{\{|y|>t\}}$. In the empirical best basis, apply hard-thresholding with threshold $t = \sqrt{\Lambda_n}$:

$$\hat{s}_i^*[\hat{\mathcal{B}}] = \eta_{\sqrt{\Lambda_n}}(y_i[\hat{\mathcal{B}}]).$$

Theorem: With probability exceeding $\pi_n = 1 - e/M_n$,

$$\|\hat{s}^* - s\|_2^2 \leq (1 - 8/\lambda)^{-1} \cdot \Lambda_n \cdot \min_{\mathcal{B} \in \mathcal{L}} E\|\hat{s}_\mathcal{B} - s\|_2^2.$$

Here the minimum is over all ideal procedures working in all bases of the library, i.e. in basis $\mathcal{B}$, $\hat{s}_\mathcal{B}$ is just $y_i[\mathcal{B}]1_{\{|s_i[\mathcal{B}]|>1\}}$.

In short, the basis-adaptive estimator achieves a loss within a logarithmic factor of the ideal risk which would be achievable if one had available an oracle which would supply perfect information about the ideal basis in which to de-noise, and also about which coordinates were large or small.

The result extends in obvious ways to more general orthogonal basis libraries, basically to any libraries constructed from an at-most polynomially-growing number of coefficient functionals. Parallel results can be developed for closely related entropies.

Key Words. Wavelet Packets, Cosine Packets, weak-ℓ^p spaces. Adaptive Basis Selection. Oracles for adaptation. Thresholding of Wavelet Coefficients.

1 Introduction

Suppose we have noisy data

$$y_i = s_i + z_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $s = (s_i)$ is the signal of interest and the (z_i) are i.i.d. $N(0,1)$. (The case of more general noise variance is handled by rescaling). We wish to recover s with small risk $R(\hat{s}, s) = E\|\hat{s} - s\|_2^2$. It is currently popular to use a fixed orthonormal bases such as wavelet bases for noise removal, by the following thresholding scheme: transform into the basis, apply thresholding, return to the original basis [5, 11, 10]. It has been shown that the success of such a de-noising scheme is directly tied to the extent to which the orthonormal basis compresses the signal to be recovered [6]. Since a given signal may be compressed well in one basis and not in others, one naturally expects that working in a single, fixed basis (Fourier, Wavelet, etc.) will impose limitations on the kind of signals which can be adequately de-noised.

In a series of pathbreaking papers, Coifman, Meyer, Wickerhauser and collaborators [2, 3] have developed the powerful ideas of libraries of *rapidly constructible* orthonormal bases, libraries which are *rapidly searchable* for "best bases" for representing a signal. Here "best" means most compressed according to some measure of "entropy". Owing to this, it is natural to explore the possibility of using the above adaptive basis ideas for the purpose of noise removal; Coifman and Wickerhauser have informally pursued such ideas; for formal proposals see [7, 13].

In this paper we show that a certain method for empirically selecting a basis in which to adaptively denoise attains near-ideal performance, in a precise sense.

2 Ideal De-noising in a fixed Basis

We summarize ideas discussed at greater length in [8]. Suppose we have noisy data as in (1). With the suffix $[\mathcal{B}]$ on a vector denoting the transform of the indicated vector into orthonormal basis $\mathcal{B}$, we have, by Parseval, that $z[\mathcal{B}]$ is a white noise and that $\|s'[\mathcal{B}] - s[\mathcal{B}]\|_2 = \|s' - s\|_2$ where the absence of a suffix denotes the use of the natural basis.

Suppose we have an estimator $\hat{s}(y)$, its mean-squared error or *risk* is $R(\hat{s}, s) = E\|\hat{s}(y) - s\|_2^2$. Consider now a specific type of estimator: an estimator which is a diagonal projector in basis $\mathcal{B}$. Then $\hat{s}(y)[\mathcal{B}] = w_i y_i[\mathcal{B}]$, where the w_i are either 1 or 0. The squared-error in a coordinate where $w_i = 1$ is the noise variance 1; the squared error in a coordinate where $w_i = 0$ is the coordinate energy $(s_i[\mathcal{B}])^2$ The minimum mean-squared error is thus $\min(s_i[\mathcal{B}]^2, 1)$ and is achieved by the choice of constants $w_i^* = 1_{\{|s_i[\mathcal{B}]|>1\}}$.

As the optimal constants w_i^* depend on the unknown signal, they are not generally available to us, unless we have available an *oracle*. The risk we can achieve with such an oracle in basis $\mathcal{B}$ is

$$\mathcal{R}(s, \mathcal{B}) = \sum_{i=1}^{n} \min(s_i[\mathcal{B}]^2, 1).$$

This is an ideal risk, attainable only by an ideal procedure (the side information w_i^* provided by an oracle is necessary). (We could also consider allowing diagonal linear estimators; i.e.

estimators which are diagonal linear in basis $\mathcal{B}$ employing arbitrary weights w_i. However, the gain in allowing general weights is not large, as the two ideal risks differ by at most a factor 2.)

Donoho and Johnstone [8] have shown that it is possible to come within a factor $2\log(n)$ of these ideal risks. Let $\lambda_n = 2\log(n)$ and define the hard thresholding nonlinearity $\eta_t(y) = y1_{\{|y|>t\}}$. Set

$$\hat{s}_i[\mathcal{B}] = \eta_{\sqrt{\lambda_n}}(y_i[\mathcal{B}]);$$

one obtains

$$R(\hat{s}, s) \leq (\lambda_n + 2.4) \cdot (1 + \mathcal{R}(s, \mathcal{B})), n = 4, 5, 6, \ldots, \tag{2}$$

and that no essentially better inequality can hold universally, for all signals $s \in R^n$.

3 Ideal De-Noising in an adaptively chosen basis

The preceding notions extend immediately to the case where one has a library $\mathcal{L}$ consisting of finitely many orthonormal bases $\mathcal{L} = \{\mathcal{B}_1, \ldots, \mathcal{B}_L\}$.

The best ideal risk in any basis in the library is

$$\mathcal{R}^*(s, \mathcal{L}) = \min_{\mathcal{B} \in \mathcal{L}} \mathcal{R}(s, \mathcal{B});$$

of course this risk is achievable only with the aid of a *basis* oracle, which selects for us the basis achieving the optimum; a *coordinate* oracle informing which coordinates in that basis are worth estimating is also necessary.

To explain the benefit of denoising in an ideal basis, we mention a connection with data compression. Let $(|s[\mathcal{B}]|_{(i)})$ denote the decreasing rearrangement of the coefficients in $s[\mathcal{B}]$, so that $|s[\mathcal{B}]|_{(1)} = \max_i |s_i[\mathcal{B}]|$ and $|s[\mathcal{B}]|_{(n)} = \min_i |s_i[\mathcal{B}]|$. Let $C(s[\mathcal{B}], m)$ be the compression number $\sum_{k>m} |s[\mathcal{B}]|_{(i)}^2$ measuring the error of reconstruction of s from its m-largest terms in an expansion in basis $\mathcal{B}$. Let $N_{\mathcal{B}}(s, \epsilon)$ denote the number of coordinates of $s_i[\mathcal{B}]$ in basis $\mathcal{B}$ exceeding ϵ in absolute value. Put for short $N_{\mathcal{B}} = N_{\mathcal{B}}(s, 1)$. Then we have a relation between ideal risk and compression numbers:

$$\mathcal{R}(s, \mathcal{B}) = C(s[\mathcal{B}], N_{\mathcal{B}}) + N_{\mathcal{B}}.$$

If an object has a very compressed representation in basis $\mathcal{B}$ then both $C(s[\mathcal{B}], N_{\mathcal{B}})$ and $N_{\mathcal{B}}$ will be small, and so the ideal risk in that basis will be small. Moreover, if the object has a small ideal risk in basis $\mathcal{B}$, the compression numbers will likewise be small.

Consider now the case where the library $\mathcal{L}$ consist of two bases, the Fourier and the Haar basis. Let the signal $s^{(1)}$ be a discretization of a Heaviside: $s_i^{(1)} = 1_{\{i>t_0 \cdot n\}}$ with $t_0 \in (0, 1)$. Then using the above notions we can calculate

$$\mathcal{R}(s^{(1)}, Fourier) \geq c \cdot \sqrt{n}, \qquad \mathcal{R}(s^{(1)}, Haar) \leq 2\log_2(n).$$

On the other hand, let $s^{(2)}$ be a smoothly localized sinusoid at the Nyquist frequency: $s_i^{(2)} = \exp(-(i - n/2)^2/(10n^2)) \sin(\pi i)$. Then

$$\mathcal{R}(s^{(2)}, Fourier) \leq C \cdot \log(n), \qquad \mathcal{R}(s^{(2)}, Haar) \geq c \cdot n.$$

In both cases we have that working in an ideally-selected basis

$$\max_{j=1,2} \mathcal{R}^*(s^{(j)}, \mathcal{L}) \leq C \log(n),$$

while working in a fixed basis fares far worse:

$$\max_{j=1,2} \mathcal{R}(s^{(j)}, Fourier) \geq c \cdot \sqrt{n}, \qquad \max_{j=1,2} \mathcal{R}(s^{(j)}, Haar) \geq c \cdot n.$$

It would, obviously, be desirable to have an algorithm that selects a basis in a near-ideal fashion. However, there seems reason to doubt that this is possible. Note that in the case of the complete wavelet packet library for signals of length n, the cardinality of library $\mathcal{L} > 2^n$. It seems plausible that if the number L of orthonormal bases in the library $\mathcal{L}$ is very large, then it will be very difficult to mimic a basis oracle, since the task of searching through a very large number of bases will lead to inevitable mistakes – bases which apparently look good, but only because of noise fluctuations and over-zealous data mining.

As it turns out, using the right entropy, one can select, empirically, a basis with near-ideal properties. We suppose that the library has the following structure: the collection of all L_n bases in the library $\mathcal{L}$ contains M_n different vectors. For example, suppose we are working with a complete wavelet-packet library. The complete table of all coefficient functionals in all wavelet packet bases has $M_n = n \log_2(n)$ coefficients. Let $t_n = \sqrt{2\log(M_n)}$, let $\lambda > 8$, and let $\Lambda_n = (\lambda \cdot (1 + t_n))^2$. Note that Λ_n is not materially larger than $2\log(n)$.

Define the empirical entropy

$$\mathcal{E}_\lambda(y, \mathcal{B}) = \sum_i \min(y_i^2[\mathcal{B}], \Lambda_n).$$

Let $\hat{\mathcal{B}}$ be the best orthonormal basis according to this entropy:

$$\hat{\mathcal{B}} = \arg\min_{\mathcal{B} \in \mathcal{L}} \mathcal{E}_\lambda(y, \mathcal{B}).$$

In the empirical best basis, apply the hard-threshold de-noising

$$\hat{s}_i^*[\hat{\mathcal{B}}] = \eta_{\sqrt{\Lambda_n}}(y_i[\hat{\mathcal{B}}]).$$

Theorem 1 *With probability exceeding $\pi_n = 1 - e/M_n$,*

$$\|\hat{s}^* - s\|_2^2 \leq (1 - 8/\lambda)^{-1} \cdot \Lambda_n \cdot \mathcal{R}^*(s, \mathcal{L}). \tag{3}$$

In the above result, it happens that the size of the library enters not through the number of bases, which may, as in the wavelet packets library, be exponentially large, but through the log of the number M_n of vectors in the library; and that results differ from ideal by only a logarithmic factor in n provided that the number of distinct vectors is at most polynomial in n.

In the normalization we have chosen, the ideal risk $\mathcal{R}^*(s, \mathcal{L})$ is typically of size a power of n for usual signals; like n for vectors which are poorly compressed in any basis, and like n^{1-r} for an $r \in (0, 1)$, which are well compressed in some basis. Hence when M_n grows at most polynomially the term Λ_n grows logarithmically, and does not typically change the rate behaviors.

Comparing (3) with (2), one sees that the result is very nearly the same as in the fixed basis case. Suppose that the number M_n grows at most polynomially in n. Then, in the ideal basis, the RHS of (3) is larger than the RHS of (2) by at most a constant factor. Thus selecting a basis worsens the risk upper bound by at most a constant factor, compared to working in an ideal basis. Since (2) has been shown in [8] to be in some sense optimal, clearly (3) cannot be significantly strengthened either.

4 Proof by Minimum Complexity Argument

Define the *Complexity* functional

$$K(\tilde{s}, s) = \|\tilde{s} - s\|_2^2 + \Lambda_n N_{\mathcal{L}}(\tilde{s}),$$

where

$$N_{\mathcal{L}}(\tilde{s}) = \min_{\mathcal{B}} \#\{i : \tilde{s}_i[\mathcal{B}] \neq 0\}$$

is the minimum complexity (non-zeroness) of $\tilde{s}$ in any basis in the library $\mathcal{L}$.

We make three simple observations, which the reader should verify:

K1. The estimator $\hat{s}^*$ is the empirical minimum complexity estimate:

$$\hat{s}^* = \arg\,\min K(\tilde{s}, y).$$

K2. The theoretical complexity of $\hat{s}^*$ upperbounds the loss:

$$K(\hat{s}^*, s) \geq \|\hat{s}^* - s\|_2^2.$$

K3. The minimum theoretical complexity is within a logarithmic factor of the ideal risk:

$$\begin{aligned} \min_{\tilde{s}} K(\tilde{s}, s) &= \min_{\mathcal{B}} \sum_i \min(s_i^2[\mathcal{B}], \Lambda_n) \\ &\leq \Lambda_n \cdot \min_{\mathcal{B}} \sum_i \min(s_i^2[\mathcal{B}], 1) \\ &= \Lambda_n \cdot \mathcal{R}(\mathcal{L}, s). \end{aligned}$$

In view of the above substititions, our main result reduces to the following bound, which says that the empirical minimum complexity estimate achieves almost theoretical minimum complexity.

Complexity Bound. *With probability at least* $\pi_n = 1 - e/M_n$,

$$K(\hat{s}^*, s) \leq (1 - 8/\lambda)^{-1} \cdot \min_{\tilde{s}} K(\tilde{s}, s). \tag{4}$$

We first let s^0 denote a signal of minimum theoretical complexity:

$$K(s^0, s) = \min_{\tilde{s}} K(\tilde{s}, s).$$

As $\hat{s}^*$ has minimum empirical complexity,

$$K(\hat{s}^*, y) \leq K(s^0, y).$$

As $\|\hat{s}^* - y\|_2^2 = \|\hat{s}^* - s - z\|_2^2$ we can relate empirical and theoretical complexities by

$$K(\hat{s}^*, y) = K(\hat{s}^*, s) + 2\langle z, s - \hat{s}^* \rangle + \|z\|_2^2,$$

and so combining the last two displays,

$$K(\hat{s}^*, s) \leq K(s^0, s) + 2\langle z, \hat{s}^* - s^0 \rangle.$$

Now define the random variable

$$\mathcal{W}(k) = \sup\{\langle z, s^2 - s^1 \rangle : \|s^j - s\|_2^2 \leq k, \Lambda_n N_{\mathcal{L}}(s^j) \leq k\}.$$

Then $K(\hat{s}^*, s) \leq K(s^0, s) + 2\mathcal{W}(K(\hat{s}^*, s))$. Lemma 2 below bounds $\mathcal{W}(k) \leq 4k/\lambda$ for all k with probability exceeding π_n. Hence, with probability exceeding π_n,

$$K(\hat{s}^*, s) \leq K(s^0, s) + 8/\lambda \cdot K(\hat{s}^*, s).$$

(4) follows by simple algebra.

Lemma 2 *With probability exceeding π_n,*

$$\mathcal{W}(k) \leq 4k/\lambda, \qquad \text{for all } k.$$

Proof. Fix a positive integer j and consider k in the range $[j\Lambda_n, (j+1)\Lambda_n)$. (Note that $\mathcal{W}(k) \equiv 0$ for $k < \Lambda_n$.) Let s^1 and s^2 be "sparse" vectors feasible for the optimization problem $\mathcal{W}(k)$, containing at most $j = \lfloor k/\Lambda_n \rfloor$ nonzero coefficients in bases $\mathcal{B}_1$ and $\mathcal{B}_2$, say. Let $\mathcal{S} = \mathcal{S}(\mathbf{m}_1 \cup \mathbf{m}_2)$ denote the linear span of the $\leq 2j$ distinct vectors occurring nontrivially (in sets $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively) in the representations of s^1 and s^2. Then let $P_{\mathcal{S}}$ denote orthonormal projection onto $\mathcal{S}$. Evidently

$$|\langle z, s^2 - s^1 \rangle| \leq \|P_{\mathcal{S}} z\|_2 \cdot \|s^2 - s^1\| \leq 2\sqrt{k}\|P_{\mathcal{S}} z\|_2, \qquad (5)$$

where $\|s^2 - s^1\| \leq 2\sqrt{k}$ because the s^j are feasible for the optimization problem.

Let $\mathcal{M}$ denote the collection of all distinct vectors occurring in some basis in the library $\mathcal{L}$. Let $\beta > 0$ be a constant to be determined later. On the event

$$A_j = \{\|P_{\mathcal{S}(M)} z\|_2 \leq \sqrt{2j(1+\beta)}\,(1 + \sqrt{2\log(M)}), \forall \mathbf{m} \subset \mathcal{M}, \#\mathbf{m} = 2j\}$$

we have from (5),

$$\begin{aligned} |\langle z, s^2 - s^1 \rangle| &\leq 2\sqrt{k}\sqrt{2j(1+\beta)}(1 + \sqrt{2\log(M)}) \\ &= 4\sqrt{k}\sqrt{j\Lambda_n}/\lambda \leq 4k/\lambda \end{aligned}$$

on setting $\beta = 1$. On the event $E_n = \cap_{j \geq 1} A_j$, the bound desired by the lemma holds. Appealing to Lemma 3 below,

$$P(A_j^c) \leq 2/(M_n\,(2j)!).$$

¿From $2\sum_1^\infty 1/(2j)! \leq e$ we get $P(E_n^c) \leq e/M_n$ and Lemma 2 follows.

Lemma 3 *Let M vectors in R^n be given, and let $C(D,M)$ denote the collection of all subsets consisting of D out of those M vectors. Then for $\beta > 0$*

$$P(\sup_{\mathcal{S}\in C(D,M)} \|P_{\mathcal{S}} z\|_2 > \sqrt{D}(1+\sqrt{2(1+\beta)\log(M)})) \leq 2M^{-\beta}/D!. \tag{6}$$

Proof. As $\|P_{\mathcal{S}} z\|_2$ is a Lipschitz functional on Gauss space, with $E\|P_{\mathcal{S}} z\|_2^2 = D$, Borell's inequality [14] gives

$$P(\|P_{\mathcal{S}} z\|_2 > \sqrt{D} + t) \leq 2e^{-t^2/2} \qquad t > 0.$$

Setting $t = \sqrt{D} \cdot \sqrt{2(1+\beta)\log(M)}$, we have $e^{-t^2/2} = M^{-D(1+\beta)}$. Letting p denote the probability in the left-hand side of (6),

$$p \leq \#C(D,M) \cdot 2e^{-t^2/2} = \binom{M}{D} \cdot 2 \cdot M^{-D(1+\beta)} \leq 2/D! \cdot M^{-\beta}.$$

5 Discussion

1. *Variations.* Other entropies and thresholding schemes admit of similar treatment. For example, the Stein's Unbiased estimate of risk for soft thresholding mentioned in [7] or the Minimum Description Length entropy proposed in [13]. Details for other entropies and thresholding schemes will appear in [9]

2. *Improved results.* A version of our main result dealing with expectations is also possible. Also, working harder, one can get results for $\lambda < 8$.

3. *Statistical Literature.* There is currently a great deal of interest in the statistical literature in convergence properties of the method of Sieves, in the abstract method of maximum likelihood, and in the method of maximum penalized likelihood. Key papers in this literature include Nemirovskii, Tsybakov and Polyak (1985), van de Geer (1988), and Birgé and Massart (1992). The argument we have given above is similar to arguments appearing in that literature; we have written our proof to emphasize the similarity. The statistical literature fcuses, generally, on a single basis; they select a best model (subset of terms in that one basis), whereas here we work in an exponentially growing collection of bases, and select both a basis, and a model in that basis. Adapting forthcoming results of Birgé and Massart on penalized estimators should allow one to obtain parallel results in density estimation, for example, and in random-design nonparametric regression models.

Acknowledgements. This work was supported by NSF DMS 92-09130, NIH CA 59039-18 (Stanford) and by ONR Contract ONR-N00014-92-0066 (Statistical Science, Inc.). We are happy to acknowledge several conversations with R.R. Coifman, with Lucien Birgé and Pascal Massart.

References

[1] L. Birgé and P. Massart, Rates of Convergence of Minimum Contrast Estimators. *Prob. Thry. and Rel. Flds.*, 1993.

[2] R. R. Coifman and Y. Meyer, “Remarques sur l’analyse de Fourier à fenêtre”, *Comptes Rendus Acad. Sci. Paris* (A) **312** (1991) p. 259-261.

[3] R. R. Coifman, Y. Meyer and M. V. Wickerhauser, “Wavelet analysis and signal processing”, pp. 153–178 in *Wavelets and Their Applications*, M. B. Ruskai et al. (eds.), Jones and Bartlett (Boston) 1992.

[4] R. R. Coifman and M. V. Wickerhauser, “Entropy-based algorithms for best-basis selection”, *IEEE Trans. Info. Theory* **38** (1992) p. 713-718.

[5] R. A. DeVore and B. J. Lucier, “Fast wavelet techniques for near-optimal image processing”, *Proc. IEEE Mil. Commun. Conf.*, Oct. 1992. IEEE, NY, 1992.

[6] D. L. Donoho, Unconditional bases are optimal bases for data compression and for statistical estimation, *Appl. Comput. Harmon. Anal.*, **1** (1993) , p. 100-115.

[7] D. L. Donoho, Nonlinear Wavelet Methods for Recovery of Signals, Images, and Densities from noisy and incomplete data, in *Different Perspectives on Wavelets*, I. Daubechies, ed. AMS: Providence, 1993.

[8] D.L. Donoho and I.M. Johnstone, Ideal Spatial Adaptation via Wavelet Shrinkage, to appear *Biometrika.*

[9] D.L. Donoho and I.M. Johnstone, *Ideal Time-Frequency Denoising.* Technical Report, Department of Statistics, Stanford University.

[10] D.L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, “Wavelet Shrinkage: Asymptopia”, to appear, *J. Roy. Statist. Soc.*.

[11] I. M. Johnstone, G. Kerkyacharian, and D. Picard, “Estimation d’une densité de probabilité par méthode d’ondelettes”, *Comptes Rendus Acad. Sciences Paris* (A) **315** (1992), p. 211–216.

[12] A.S. Nemirovskii, B.T. Polyak, and A.B. Tsybakov, Rate of convergence of nonparametric estimates of maximum-likelihood type, *Problems of Information Transmission*, **21** (1985), p. 258–272.

[13] N. Saito, Simultaneous Noise Suppression and Signal Compression using a library of orthonormal bases and the minimum description length criterion. To appear, *Wavelets in Geophysics*, Academic Press.

[14] M. Talagrand and M. Ledoux, *Probability in Banach Spaces*, Springer.

[15] S. Van de Geer, A new approach to least-squares estimation, with applications, *Ann. Statist.* **15** (1988), p. 587–602.