Shrinkage and Denoising by Minimum Message Length

Daniel F. Schmidt and Enes Makalic

***Abstract*—This paper examines orthonormal regression and wavelet denoising within the Minimum Message Length (MML) framework. A criterion for hard thresholding that naturally incorporates parameter shrinkage is derived from a hierarchical Bayes approach. Both parameters and hyperparameters are jointly estimated from the data directly by minimisation of a two-part message length, and the threshold implied by the new criterion is shown to have good asymptotic optimality properties with respect to zero-one loss under certain conditions. Empirical comparisons made against similar criteria derived from the Minimum Description Length principle demonstrate that the MML procedure is competitive in terms of squared-error loss.**

I. Introduction

Consider an observed dataset $\mathbf{y} \subset \mathbb{R}^n$ generated from the model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the *true* underlying signal, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is additive noise due to measurement error. The aim of a *denoiser* is to produce an estimate $\hat{\boldsymbol{\mu}}(\mathbf{y})$ of $\boldsymbol{\mu}$ from the noisy measurement $\mathbf{y}$ with small error, typically measured in terms of squared error loss $L(\hat{\boldsymbol{\mu}}(\mathbf{y}), \boldsymbol{\mu}) = ||\hat{\boldsymbol{\mu}}(\mathbf{y}) - \boldsymbol{\mu}||_2$. *Linear-Gaussian* denoisers make two assumptions; first that the error disturbances are normally distributed, i.e. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \tau \mathbf{I}_n)$, and second, that the true signal $\boldsymbol{\mu}$ is represented by a linear combination of orthonormal basis, say $\mathbf{X} \in \mathbb{R}^{(n \times n)}$, i.e.

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

so that the estimate $\hat{\boldsymbol{\mu}}(\mathbf{y})$ is

$$\hat{\boldsymbol{\mu}}(\mathbf{y}) = \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y})$$

where $\hat{\boldsymbol{\beta}}(\mathbf{y})$ are estimates of $\boldsymbol{\beta}$. One of the most important bases for function approximation are those built from wavelets; these allow spatial adaptivity and can be rapidly computed by the discrete wavelet transform [1].

The application of wavelets to function estimation has become an important area of research since the seminal work of Donoho, Johnstone and collaborators [2]. They derived a simple but powerful threshold for selection of the number of non-zero coefficients, and use it in the context of both hard and soft thresholding procedures. Interestingly, the chosen threshold is near minimax for bowl shaped loss functions over a wide range of function spaces. This work has inspired a range of alternative wavelet denoising procedures, including those based on Stein's unbiased estimate of risk [3], cross-validation [4], and Bayesian techniques [5], [6].

Of particular interest in the context of this paper is the denoising solution obtained through the application of the Normalised Maximum Likelihood (NML) universal model [7] to the linear-Gaussian denoising problem [8], which results in a criterion that Rissanen claims 'cannot be found nor matched by Bayesian nor orthodox statistical techniques' [9]. One aim of this article is to show that, by the application of a relatively recent result [10], a criterion very close to that obtained by Rissanen can be derived from a hierarchical Bayes point of view by application of the Minimum Message Length principle [11], [12]. Furthermore, the two-part coding nature of the criterion provides a sound basis for shrinkage of the estimates of $\boldsymbol{\beta}$ in addition to selection of the number of non-zero ('significant') coefficients.

Daniel F. Schmidt and Enes Makalic are with Monash University
Clayton School of Information Technology
Clayton Campus Victoria 3800, Australia.
Telephone:+61 3 9905 9555, Fax: +61 3 9905 9422
Email: {Daniel.Schmidt,Enes.Makalic}@infotech.monash.edu.au

II. Minimum Message Length Inference

Inference within the Minimum Message Length principle [11], [13], [12] involves searching for the fully specified model $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that results in the briefest joint encoding of the data $\mathbf{y}$ and the chosen model. The data is always encoded in two parts; the first part $I(\boldsymbol{\theta})$, or *assertion*, states the model to optimal precision, and the second part $I(\mathbf{y}|\boldsymbol{\theta})$, or detail, states the data given the chosen model. Although such two part messages contain inherit redundancies, they allow parameter estimation and model selection to be performed within the same framework. This is in contrast to the well known Minimum Description Length principle [7], [14] which generally focusses on model class selection. While a large range of approximations and message length formulae exist, the most popular is the Wallace-Freeman MML87 approximation [15]. Within this framework the joint message length $I_{87}(\mathbf{y}, \boldsymbol{\theta})$ of model $\boldsymbol{\theta}$ and data $\mathbf{y}$ is:

$$\underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{p}{2}\log \kappa_p}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{p}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})} \tag{2}$$

where $\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the Fisher information matrix, κ_p is the normalised second moment of an optimal quantising lattice in p-dimensions [16] and $\pi(\cdot)$ is a prior distribution over the parameter space. For the purposes of this paper, the message length is measured in *nits*, where one nit is equal to $\log_2 e$ bits. Inference is then performed by selecting the model $\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$ that minimises the message length expression (2). The MML87 estimator is invariant under one-to-one reparameterisations of the parameter space.

Given the subjective Bayesian nature of MML, past applications of MML87 have generally required that all hyperparameters be specified before any data is observed. This is particularly true of previous message length formulations of the linear regression problem [17], [18]. Recently a generalisation of MML87 to allow the estimation of hyperparameters has been proposed and applied to the inference of the multiple-means problem [10]. Given a hierarchical model of the form

$$\begin{aligned} \mathbf{y} &\sim p(\boldsymbol{\theta}) \\ \boldsymbol{\theta} &\sim \pi_{\boldsymbol{\theta}}(\boldsymbol{\alpha}) \\ \boldsymbol{\alpha} &\sim \pi_{\boldsymbol{\alpha}} \end{aligned}$$

where $\boldsymbol{\alpha}$ are hyperparameters and $\pi_{\boldsymbol{\alpha}}(\cdot)$ are priors for the hyperparameters, one first finds the message length of $\mathbf{y}$ given $\boldsymbol{\theta}$ conditioned on $\boldsymbol{\alpha}$

$$l(\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \frac{1}{2}\log|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + o(1) \tag{3}$$

where $l(\boldsymbol{\theta}) = -\log p(\mathbf{y}|\boldsymbol{\theta})$. To estimate the hyperparameters, one replaces $\boldsymbol{\theta}$ with the estimates $\hat{\boldsymbol{\theta}}_{87}(\mathbf{x}|\alpha)$ that minimise (3), forming a 'profile message length'. In most cases, it appears necessary to apply the curved prior correction (see [12], p. 236–237) to the MML87 message length formula; for conjugate priors this may be done in a fashion that preserves invariance of the estimators. The precision to which $\boldsymbol{\alpha}$ must be stated is determined by finding the Fisher Information of the hyperparameters

$$\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \mathrm{E}_r\left[\frac{\partial^2 I_{87}(\mathbf{y}, \hat{\boldsymbol{\theta}}_{87}(\mathbf{y}|\alpha)|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}\right]$$

where the expectation is taken with respect to the marginal distribution $r(\mathbf{y}|\boldsymbol{\alpha}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. The complete joint message length $I_{87}(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha})$ for $\mathbf{y}$, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ is then given by

$$\begin{aligned} & I_{87}(\boldsymbol{\alpha}) + I_{87}(\boldsymbol{\theta}|\boldsymbol{\alpha}) + I_{87}(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}) \\ = \;& l(\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) + \frac{1}{2}\log|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})||\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| + o(1) \end{aligned} \tag{4}$$

Finally, one may minimise $I_{87}(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\alpha})$ to jointly estimate both parameters $\boldsymbol{\theta}$ and the hyperparameters $\boldsymbol{\alpha}$. This generalised MML87 criterion is now applied to the linear-Gaussian denoising problem.

III. MML Orthonormal Regression Criterion

Consider the linear-Gaussian denoiser specified by the following hierarchy:

$$\mathbf{y} \sim N_n\left(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \tau \mathbf{I}_n\right) \quad (5)$$

$$\boldsymbol{\beta} \sim N_p\left(\mathbf{0}_p, m\mathbf{I}_p\right) \quad (6)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients, $\tau \in \mathbb{R}^+$ is the noise variance, $m \in \mathbb{R}^+$ is an unknown hyperparameter, γ is an index set and $\mathbf{X} \in \mathbb{R}^{(n \times p)}$ is an orthonormal design matrix. The aim is estimate the parameters $\boldsymbol{\beta}$, τ and m in addition to finding the best subset γ of $\mathbf{X}$, i.e. determining which regression coefficients differ significantly from zero. In the remainder of the article the explicit dependence of $\mathbf{X}$ and $\boldsymbol{\beta}$ on γ is dropped for clarity. The negative log-likelihood function conditioned on τ is:

$$l(\boldsymbol{\beta}|\tau) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\tau + \frac{1}{2\tau}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \quad (7)$$

Application of the MML87 approximation (2) requires the Fisher Information matrix $\mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ to determine the precision to which the parameters are stated. Due to the fact that the m is estimated from the data, the Fisher Information may provide a poor approximation to the coding quantum as $m \to 0$; as in [12], p. 236–237, this is corrected by forming a modified Fisher Information by treating the parameters $\boldsymbol{\beta}$ as p instances of 'prior data' with mean zero and variance $1/m$ [19]. This correction preserves the invariance property of the MML87 estimator, and yields a 'corrected' Fisher Information

$$|\mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\beta})| = \left(\frac{1}{\tau} + \frac{1}{m}\right)^p \quad (8)$$

Rather than jointly estimating the hyperparameters $\boldsymbol{\alpha} = (m, \tau)$ and $\boldsymbol{\beta}$ as in [20], $\boldsymbol{\beta}$ is estimated conditioned on $\boldsymbol{\alpha}$, and then τ and m are jointly estimated. Substituting (6), (7) and (8) into (2) and minimising for $\boldsymbol{\beta}$ yields the MML87 estimator

$$\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}|\boldsymbol{\alpha}) = \left(\frac{m}{m+\tau}\right)\mathbf{X}'\mathbf{y} = \left(\frac{m}{m+\tau}\right)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\mathbf{y}) \quad (9)$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\mathbf{y})$ is the usual least-squares estimator. This is clearly a James-Stein type shrinkage estimator where the amount of shrinkage depends on the variance m of the regression parameters as well as the noise variance τ. As $\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}|\boldsymbol{\alpha})$ clearly depends on m and τ the message length format is reordered to transmit the hyperparameters before the regression parameters. It remains to estimate τ and m; as $\hat{\boldsymbol{\beta}}_{87}(\cdot)$ is conditioned on both of these quantities the precision to which they must be stated depends on the 'profile message length' formed by replacing $\boldsymbol{\beta}$ with the MML87 estimate. Following the procedure in [10] and after some simplifications the profile message length $I_{87}(\mathbf{y}, \hat{\boldsymbol{\beta}}_{87}(\mathbf{y}|m)|\tau, m)$ is given by

$$\begin{aligned} &\frac{n+p}{2}\log(2\pi) + \frac{n}{2}\log\tau + \frac{p}{2}\log m + \frac{m\xi(\mathbf{y})}{2(m+\tau)^2} \\ &+\frac{p}{2}\log\left(\frac{1}{\tau}+\frac{1}{m}\right) + \frac{1}{2\tau}\left(\mathbf{y}'\mathbf{y} - \frac{2m\xi(\mathbf{y})}{m+\tau} + \frac{m^2\xi(\mathbf{y})}{(m+\tau)^2}\right) \\ &+\frac{p}{2}\left(\log\kappa_p + 1\right) \end{aligned} \quad (10)$$

where $\xi(\mathbf{y}) = \mathbf{y}'(\mathbf{X}\mathbf{X}')\mathbf{y} = ||\hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\mathbf{y})||_2^2$ is the fitted sum-of-squares.

When computing the message length of hierarchical Bayes models the matrix $\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ of expected second-derivatives of m and τ is used to determine precision to which these hyperparameters must be stated; the expectations are taken with respect to the marginal distribution of $\mathbf{y}$ conditioned on m and τ. Noting that $\mathrm{E}\left[\xi(\mathbf{y})\right] = p(m+\tau)$ and $\mathrm{E}\left[\mathbf{y}'\mathbf{y}\right] = \mathrm{E}\left[\xi(\mathbf{y})\right] + (n-p)\tau$ the Fisher Information of $\boldsymbol{\alpha}$ is:

$$|\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| = \frac{(n-p)p}{4\tau^2(m+\tau)^2} \quad (11)$$

It is clear that the smaller m and τ become the smaller the coding quantum for $\boldsymbol{\alpha}$. Finally, priors for the hyperparameters must be specified; the prior

$$\pi(\boldsymbol{\alpha}) \propto \tau^{-\nu} \quad (12)$$

over some suitable range is chosen as suitably uninformative, with ν a user selected hyperparameter. Several choices for ν are discussed in the sequel. Using (6), (7), (8), (11) and (12) in (4) yields the following MML estimators for τ and m

$$\hat{\tau}_{87}(\mathbf{y}) = \frac{\mathbf{y}'\mathbf{y} - \xi(\mathbf{y})}{n - p + 2\nu - 2} \quad (13)$$

$$\hat{m}_{87}(\mathbf{y}) = \left(\frac{\xi(\mathbf{y}) - \hat{\tau}_{87}(\mathbf{y})(p-2)}{p-2}\right)_+ \quad (14)$$

where $(\cdot)_+ = \max(\cdot, 0)$ as m may never be negative.

To state the set γ of retained regressors a Bernoulli prior is placed over the indicator for each coefficient, yielding a code length

$$I_{87}(\gamma) = -p\log\eta - (n-p)\log(1-\eta) \quad (15)$$

where η is a hyperparameter chosen to reflect the *a priori* expected number of non-zero coefficients. Using the approximation (p. 237, [12])

$$\frac{p}{2}\left(\log\kappa_p + 1\right) \approx -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log(p\pi) + \psi(1) \quad (16)$$

where $\psi(\cdot)$ is the digamma function, the complete joint message length, including the dependency on γ, is

$$I_{87}(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma) = I_{87}(\gamma) + I_{87}(\boldsymbol{\alpha}|\gamma) + I_{87}(\boldsymbol{\beta}|\boldsymbol{\alpha}, \gamma) + I_{87}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma)$$

which simplifies to

$$\begin{aligned} I_{87}(\mathbf{y}, \boldsymbol{\theta}) = &\frac{n-p-2\nu+2}{2}\log\hat{\tau}_{87}(\mathbf{y}) + \frac{p-2}{2}\log\frac{\xi(\mathbf{y})}{p-2} \\ &+\frac{1}{2}\log(n-p)p^2 - p\log\eta - (n-p)\log(1-\eta) + c \end{aligned} \quad (17)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma)$,

$$c = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\frac{\pi}{4} + \log\Omega + \frac{n}{2} + \nu - 2 + \psi(1)$$

are all terms not depending on p, and Ω is a normalising constant for the improper prior $\pi(\boldsymbol{\alpha})$. The set γ that minimises (17) is considered the optimal model. Note that (17) is similar to the criterion derived by Rissanen [8] using the Normalised Maximum Likelihood (NML) model (see Section V for discussion).

Remark 1: Behaviour of $\hat{\tau}_{87}(\mathbf{y})$. The estimate $\hat{\tau}_{87}(\mathbf{y})$ is based on the residual variance of the least-squares estimates rather than the residuals $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}))$ of the shrunken MML estimates, despite these residuals representing the incompressible noise in the two-part MML message. This implies that a choice of $\nu = 1$ yields the usual unbiased estimate of variance.

Remark 2: Behaviour of the shrinkage factor. The estimate (14) is very close to the usual James-Stein positive part estimate. In particular, for $\nu = 2$ the estimate coincides with the modification proposed by Sclove [21] that is known to dominate least-squares in terms of squared error loss.

Remark 3: Behaviour for $p < 3$. When $p < 3$, it is straightforward to show that $\hat{m}_{87}(\mathbf{y}) = \infty$. This diffuse choice of hyperparameter

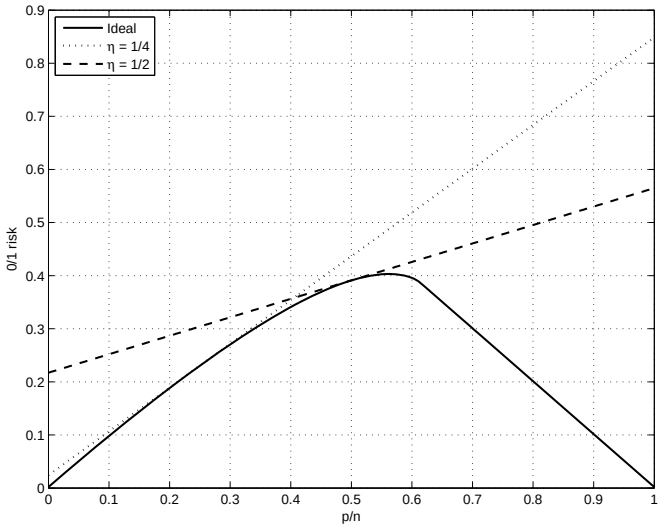


Fig. 1. 0/1 Risks for Ideal and MML Thresholding

causes the MML87 estimates to coincide with the least squares estimates, which are known to be inadmissable for $p \leq 2$. However, this also causes the message length approximation to break down and results in nonsensical message lengths which cannot be used for model selection. A straightforward 'fix' is to replace the estimate for m with the modified estimate

$$\hat{m}_{87}(\mathbf{y}) = \left(\frac{\xi(\mathbf{y}) - \hat{\tau}_{87}(\mathbf{y})(p-2)}{\max(p-2,1)} \right)_+$$

While not entirely satisfactory, this modification appears to yield reasonable message lengths for models with $p \leq 2$ with some degradation in squared error risk. This is of no great concern in the denoising setting as the number of retained regressors p is generally much greater than two.

Remark 4: Searching for the optimal model. Application of Theorem 2 in [8] shows that the set γ that minimises (17) contains the p largest (in absolute value) regression coefficients of the full regression. The search is then reduced to a search for optimal p which may be completed in $O(n)$ time.

A. MML Threshold

An approximate value of the threshold $\hat{\lambda}_{87}(\mathbf{y})$ selected by the MML orthonormal regression criterion can be derived in terms of the variance m of the non-zero coefficients of the underlying signal, and the noise variance τ. Considering the change in message length $I(\mathbf{y}, \cdot|\gamma)$ when a coefficient with variance λ^2 is added to a model and let $I(\mathbf{y}, \cdot|\gamma')$ denote the message length of a model with the expanded regressor set γ'. Assuming that p and n are sufficiently large such that the estimates $\hat{m}_{87}(\mathbf{y})$ and $\hat{\tau}_{87}(\mathbf{y})$ possess small variance and are unchanged by the inclusion of the new coefficient λ, the change in message length $I_{87}(\mathbf{y}, \cdot|\gamma) - I_{87}(\mathbf{y}, \cdot|\gamma')$ is

$$\frac{\lambda^2 m}{2(m+\tau)\tau} - \frac{1}{2}\log\left(\frac{m+\tau}{\tau}\right) + \log\left(\frac{\eta}{1-\eta}\right)$$

The MML87 criterion only includes a regressor if this difference is negative, i.e. if including the regressor reduces the message length. Solving this expression for zero yields the asymptotic threshold

$$\hat{\lambda}^2_{\infty} = \left(\frac{\tau(m+\tau)}{m}\right)\log\left(\frac{(m+\tau)(1-\eta)^2}{\tau\eta^2}\right) \qquad (18)$$

such that all coefficients with absolute value less than $\hat{\lambda}_{\infty}$ are set to zero. Empirical testing indicates that for moderate p and n, and for ratios of m/τ not too close to unity, the threshold given by (18) matches closely the threshold $\hat{\lambda}_{87}(\mathbf{y})$ chosen by the MML87 criterion. If one could obtain reliable estimates of m and τ *a priori* then the MML (and MDL) thresholding could approximately be reduced to an application of (18); of course, the estimation of m and τ by dividing the data into noise and signal bearing components is the very problem the MML criterion aims to solve.

It is interesting to examine the performance of the thresholding procedure in terms of the frequentist 0/1 risk based on the true m, τ and the true number of non-zero coefficients p. The 0/1 loss is the sum of Type I and Type II errors attained by a thresholding procedure, i.e. the number of information bearing coefficients 'misclassified' as noise plus the number of noise coefficients that escape the threshold. After noise is added to the underlying signal there are p information bearing coefficients distributed as per $N(0, m+\tau)$, and $(n-p)$ noise coefficients distributed as per $N(0,\tau)$. For a known m, τ and p the 0/1 risk is

$$R_{0/1}(m,\tau,p,\lambda) = p\Phi_{\lambda}(m+\tau) + (n-p)\left(1-\Phi_{\lambda}(\tau)\right)$$

where $\Phi_{\lambda}(\cdot)$ is the probability that an $N(0,\cdot)$ variate lies between $[-\lambda,\lambda]$. Interestingly, the value of λ that minimises the 0/1 risk coincides exactly with the MML threshold (18) when $\eta = p/n$. This means that under the aforementioned conditions no criterion can expect to do better than MML in terms of 0/1 risk. Clearly this result is contingent on a good guess of the number of signal bearing coefficients, formalised by the choice of the hyperparameter η. Given the similarity of the MDL denoising criterion to (17), this result helps explain the 'poor' performance of the MDL denoising criterion in pure noise situations. The choice of codebook made by Rissanen in [8] is equivalent to a choice of $\eta = 1/2$, which suggests an *a priori* expectation that $n/2$ coefficients are signal bearing. In the case of pure noise, i.e $p = 0$, the risk obtained by the MDL/MML threshold with $\eta = 1/2$ is much greater than the risk obtained when η is close to zero. Of course, the MML threshold will be suboptimal if the signal bearing coefficients are not normally distributed, and this supports similar findings in [22] on the behaviour of MDL denoising.

B. Shrinkage towards a data-dependent origin

A possible extension to (17) is use a non-zero prior mean for the parameters $\boldsymbol{\beta}$. There is no compelling reason to shrink towards the zero origin (see Lindley's comments in [23] and Stone's comments in [24]), and allowing the data to determine the shrinkage origin leads to better global risk properties at the expense of mildly degraded risk near $\boldsymbol{\beta} = \mathbf{0}_p$. The new hierarchy is then

$$\mathbf{y} \sim N_n\left(\mathbf{X}_{\gamma}\boldsymbol{\beta}_{\gamma}, \tau\mathbf{I}_n\right) \qquad (19)$$

$$\boldsymbol{\beta} \sim N_p\left(a\mathbf{1}_p, m\mathbf{I}_p\right) \qquad (20)$$

The MML87 estimator for $\boldsymbol{\beta}$, conditioned on $\boldsymbol{\alpha} = (m, a, \tau)$ is

$$\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}|\boldsymbol{\alpha}) = \left(\frac{m}{m+\tau}\right)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\mathbf{y}) + \left(\frac{a\tau}{m+\tau}\right)\mathbf{1}_p$$

This estimator clearly shrinks the least squares estimates towards $a\mathbf{1}_p$, the amount of shrinkage determined by the ratio m/τ. To find the message lengths for this new hierarchy we apply formula (4) as previously. Noting that $\mathrm{E}\left[\hat{\boldsymbol{\beta}}_{\mathrm{LS}}(\mathbf{y})'\mathbf{1}_p\right] = pa$, $\mathrm{E}\left[\xi(\mathbf{y})\right] = p(m+\tau) + pa^2$ and $\mathrm{E}\left[\mathbf{y}'\mathbf{y}\right] = \mathrm{E}\left[\xi(\mathbf{y})\right] + (n-p)\tau$, the Fisher Information for the hyperparameters is

$$|\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| = \frac{(n-p)p^2}{4\tau^2(m+\tau)^3} \qquad (21)$$

Method	SNR	Model selection criteria						
		$\text{MML}_{1/2}$	MML_a	$\text{MML}_{1/4}$	MML_ϕ	NML	NML_r	SureShrink
Blocks	1	430.53 (33.80)	429.89 (33.91)	183.36 (27.38)	**171.40** (22.19)	471.88 (34.35)	195.28 (22.89)	201.22 (80.78)
	10	30.270 (3.683)	30.117 (3.675)	**25.766 (2.974)**	33.448 (4.537)	31.140 (3.783)	33.244 (4.761)	32.334 (6.121)
	25	11.072 (1.270)	**11.049** (1.265)	11.602 (1.455)	15.077 (2.879)	11.192 (1.285)	14.550 (2.418)	14.021 (2.220)
	50	5.4411 (0.612)	**5.4376** (0.611)	6.1733 (0.797)	7.3349 (1.215)	5.4673 (0.614)	7.2734 (1.154)	7.2876 (1.046)
	100	**2.7881** (0.299)	2.7885 (0.299)	3.2142 (0.407)	3.6052 (0.567)	2.7919 (0.299)	3.5504 (0.536)	3.7786 (0.500)
Bumps	1	434.72 (33.23)	433.92 (33.33)	**221.38** (27.59)	288.33 (46.51)	477.07 (33.80)	324.32 (45.62)	275.67 (68.34)
	10	32.854 (3.661)	32.755 (3.639)	**28.325** (3.314)	34.318 (5.526)	33.842 (3.762)	34.433 (5.348)	36.513 (5.086)
	25	12.213 (1.267)	**12.202** (1.266)	12.480 (1.367)	13.920 (1.961)	12.362 (1.286)	13.826 (1.808)	15.373 (1.903)
	50	6.1216 (0.625)	**6.1171** (0.625)	7.0479 (0.797)	7.5922 (1.050)	6.1536 (0.628)	7.6443 (0.986)	7.8920 (0.917)
	100	3.0862 (0.325)	**3.0849** (0.325)	3.7872 (0.509)	3.8890 (0.617)	3.0908 (0.324)	3.9772 (0.608)	4.0199 (0.437)
Doppler	1	423.72 (34.83)	422.91 (35.04)	130.69 (30.12)	**83.160** (14.85)	463.76 (35.30)	91.504 (18.80)	188.36 (81.26)
	10	20.654 (4.186)	20.324 (4.176)	**12.147** (2.218)	13.321 (2.638)	21.461 (4.277)	13.303 (2.550)	25.173 (7.352)
	25	6.4108 (1.219)	6.3450 (1.213)	**5.0102** (0.885)	5.7560 (1.073)	6.5446 (1.242)	5.7428 (1.061)	11.261 (2.699)
	50	2.8817 (0.529)	2.8619 (0.526)	**2.5355** (0.431)	2.9971 (0.574)	2.9185 (0.539)	2.8735 (0.525)	5.9304 (1.301)
	100	1.3941 (0.242)	1.3889 (0.241)	**1.3340** (0.225)	1.5712 (0.282)	1.4028 (0.244)	1.5057 (0.266)	3.0927 (0.631)
Heavisine	1	417.66 (35.98)	416.86 (36.10)	76.559 (34.72)	**17.400** (4.657)	456.85 (36.59)	18.673 (6.606)	147.16 (86.06)
	10	12.864 (4.423)	12.140 (4.400)	6.0528 (1.669)	**5.0245** (0.918)	14.032 (4.500)	5.4835 (1.255)	16.306 (8.349)
	25	4.1526 (1.050)	4.0285 (1.024)	**3.1330** (0.605)	3.2794 (0.545)	4.3563 (1.092)	3.4088 (0.586)	7.3412 (3.275)
	50	2.1550 (0.430)	2.1288 (0.419)	**1.9807** (0.331)	2.2829 (0.370)	2.1964 (0.441)	2.3302 (0.394)	4.0635 (1.591)
	100	1.1973 (0.206)	**1.1940** (0.205)	1.2377 (0.198)	1.5785 (0.249)	1.2030 (0.209)	1.5465 (0.248)	2.2190 (0.791)

TABLE I
SQUARED PREDICTION ERRORS FOR DONOHO-JOHNSTONE TEST SIGNALS

The new MML87 estimators are

$$\begin{aligned}\hat{\tau}_{87}(\mathbf{y}) &= \frac{\mathbf{y}'\mathbf{y} - \xi(\mathbf{y})}{n - p + 2\nu - 2} \\ \hat{a}_{87}(\mathbf{y}) &= \frac{\hat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{y})'\mathbf{1}_p}{p} \\ \hat{m}_{87}(\mathbf{y}) &= \left(\frac{||\hat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{y}) - \hat{a}_{87}(\mathbf{y})\mathbf{1}_p||_2^2}{p-3} - \hat{\tau}_{87}(\mathbf{y})\right)_+\end{aligned}$$

It is clear that the estimate for a is simply the arithmetic mean of $\hat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{y})$, and the estimate for τ remains unchanged. Due to the fact that a is now estimated from the data, the estimate $\hat{m}_{87}(\mathbf{y})$ is infinite when $p < 4$ rather than $p < 3$; in this case, the least squares estimates cannot be dominated by estimates from a hierarchy of the form (19), (20). However, this once again causes issues with model selection as the resulting message lengths are nonsensical; the problem is addressed as in Remark 3.

IV. MML WAVELET DENOISING

This section considers the MML orthonormal regression criterion (17) in the context of wavelet denoising. The results in Section III-A suggest that if a reasonable prior is chosen for the coefficients the MML threshold will be close to optimal in terms of 0/1 risk, at least asymptotically. Therefore, applying the MML orthonormal regression criterion to wavelet denoising amounts to choosing a distribution over the set γ that reflects the behaviour of wavelet coefficients found in real signals.

A. Selection of η

The MDL denoising criterion developed by Rissanen [8] uses a 'prior' that considers every combination of retained coefficients equally likely; this prior has the property of achieving minimax coding regret and is similar to the choice of $\eta = 1/2$. As discussed in Section III-A this can lead to poor performance when the number of non-zero coefficients is significantly smaller than $n/2$. An assumption made in many wavelet denoising criteria is that at most $n/2$ of the wavelet coefficients are non-zero (for example in [2] the $n/2$ coefficients in the finest detail scale are used to estimate the noise variance). Figure 1 shows that the 0/1 risk attained by the MML/MDL criterion for the choice $\eta = 1/2$ can be considerable when $p/n < 0.3$. In contrast, if one assumes $p \leq n/2$ then the choice $\eta = 1/4$ appeared to minimise the average 0/1 risk for $p \in [0, n/2]$ irrespective of the ratio m/τ; this suggests that $\eta = 1/4$ is a more 'universal' choice in the case of function estimation by wavelet thresholding.

B. Prior on Wavelet Scales

The aforementioned choice of prior still says little about the *a priori* expected smoothness of the underlying function; alternatively one may consider a prior that expects the number of coefficients in each coefficient level to be roughly the same [5]. Practically, this implies that the prior places less probability on coefficients being non-zero further down the levels. The chosen negative log-prior of γ is

$$l(\gamma|\phi) = -\sum_{j=1}^{J}\Big(k_j \log \phi^j + (2^j - k_j)\log(1-\phi^j)\Big)$$

where ϕ^j is the *a priori* probability that a coefficient at level j is non-zero, and $J = \log_2 n$ is the number of wavelet resolutions. Rather than treat ϕ as a known hyperparameter, we choose to estimate it from the data to minimise the total message length using MML87. Noting that the expected number of non-zero coefficients at level j is $\text{E}[k_j] = 2^j\phi^j$, the Fisher Information for ϕ is

$$J_\phi(\phi) = \frac{1}{\phi^2}\sum_{j=1}^{J}\frac{j^2 2^j\left(\phi^j - \phi^{2j}\right)}{(\phi^j - 1)^2}$$

giving a total codelength

$$I_{87}(\gamma, \phi) = l(\gamma|\phi) + \frac{1}{2}\log J_\phi(\phi) + o(1) \tag{22}$$

where the $o(1)$ term can be ignored as it is present for all models indexed by γ. No closed form solution for $\hat{\phi}_{87}(\gamma)$ appears to exist and so it must be found numerically.

V. Relation to MDL Denoising

As previously mentioned, the resulting MML denoising criterion (17) found through a hierarchical Bayes approach derived in Section III is similar in form to the MDL denoising criterion

$$\mathrm{NML}(\mathbf{y}) = \frac{n-p}{2}\log\frac{\mathbf{y}'\mathbf{y}-\xi(\mathbf{y})}{n-p} + \frac{p}{2}\log\frac{\xi(\mathbf{y})}{p} + \frac{1}{2}\log(p(n-p)) \tag{23}$$

There are several differences between the two criteria; perhaps most importantly is the fact that the NML criterion appears to advocate the use of the Maximum Likelihood estimates, while in the MML approach parameter estimation is an integral of the framework. In the case of orthonormal denoising the MML approach naturally yields shrunken estimates which for sensible choices of ν dominate the Maximum Likelihood/least squares estimates for moderate values of p, and generally outperform them even for small p. In fact, for a choice of $\nu = 2$ in (12) the MML estimate of m naturally leads to a linear shrinkage estimator that is well known to dominate least squares for all $p \geq 3$. The differences between the two codelengths (17) and (23) can be mostly attributed to the fact that the MML codes are two-part and the extra length allows for the parameter estimates to be stated. Interestingly, while the NML code is based on minimising the maximum coding regret, Roos et al [25] have shown an alternative interpretation as a 'mixture model' of two Gaussian distributions, one for signal-bearing coefficients and one for noise coefficients. This is similar to the MML interpretation that assumes a Gaussian prior on retained coefficients and a Gaussian likelihood function; however, due to the Bayesian nature of MML these assumptions are explicitly made before deriving the criterion.

Alternative MDL approaches for linear regression include the work of Liang and Barron [26] and Hansen and Yu [20]. The latter criterion is similar to (23), and is based on the so called 'Bayesian' codes. These are one-part coding schemes based on the Bayes mixture of a likelihood and a 'prior' (though in the MDL framework the interpretation as a prior is dismissed). For the g-MDL approach, Hansen and Yu adopted Gaussian priors and a Gaussian likelihood, and adjust the prior by choosing a variance hyperparameter to minimise a meta two-part code. The resulting estimate for the hyperparameter is the marginal maximum likelihood estimate, and is coded using the asymptotically optimal code length rather than an optimal finite sample codelength as in Section III. Furthermore, due to the one-part nature of the Bayes codes their criterion does not yield explicit estimates of the noise variance or the regression parameters themselves, and is similar to what Wallace calls the 'non-explanation' code ([12], page 154).

In contrast to these criteria, the strength of the MML orthonormal regression criterion developed here is that it provides a unified framework from which both model class (number of signal bearing regressors) and parameters $\boldsymbol{\beta}$ and τ may be estimated.

VI. Results

A set of numerical simulations were undertaken on the four classic Donoho and Johnstone [2] test functions: 'blocks', 'bumps', 'heavisine' and 'doppler'. All exhibit spatial inhomogeneity and are designed to analog features found in real life datasets. For each dataset $n = 512$ samples were generated and Gaussian noise at signal-to-noise (SNR) levels of $(1, 10, 25, 50, 100)$ was added to the clean signals to produce $10,000$ noisy signals for each level of noise. The following criteria were then used to produce 'denoised' versions of the signals:

- $\mathrm{MML}_{1/2}$: the criterion given in Section III with $\eta = 1/2$,
- MML_a: the criterion given in Section III-B with $\eta = 1/2$,
- $\mathrm{MML}_{1/4}$: the criterion given in Section III with $\eta = 1/4$,
- MML_ϕ: the criterion given in Section III using the multilevel prior described in Section IV-B in place of the prior (15)
- NML: the MDL denoising criterion given by (23)
- NML_r: the MDL denoising criterion with the codelength of retained coefficients given by $\log\binom{n}{p}$ as per [25].
- SureShrink: Soft thresholding based on Stein's unbiased estimate of risk [3].

In all experiments the Symlet-8 wavelet basis were employed. The error between the true signals and the denoised versions of the signals produced by each criteria were measured in terms of squared-error loss. The results are shown in Table I, with standard errors given in parentheses. As the MML criteria developed in this paper are essentially 'first generation' we chose to test them against similar MDL criteria, and have explicitly avoided comparisons with the more refined versions of MDL denoising in [22] as beyond the scope of this current paper.

In all tests the $\mathrm{MML}_{1/2}$ and MML_a methods were virtually indistinguishable and slightly outperformed NML. This was primarily due to the shrinkage applied to the parameters, and is particularly evident for $\mathrm{SNR} = 1$. The MML_ϕ and NML_r methods perform quite similarly and attain significantly lower squared errors for $\mathrm{SNR} = 1$, but due to their conservative nature do not perform as well for higher SNR values. The $\mathrm{MML}_{1/4}$ criterion offers significant improvements over $\mathrm{MML}_{1/2}$/MML_a for low SNR while remaining competitive when the noise is small, and appears to offer a good trade-off between performance at the varying levels of noise. The SureShrink criterion performed similar to MML_ϕ/NML_r on 'blocks' and 'bumps', but performed significantly worse for the 'doppler' and 'heavisine' functions which are known to have very sparse wavelet representations.

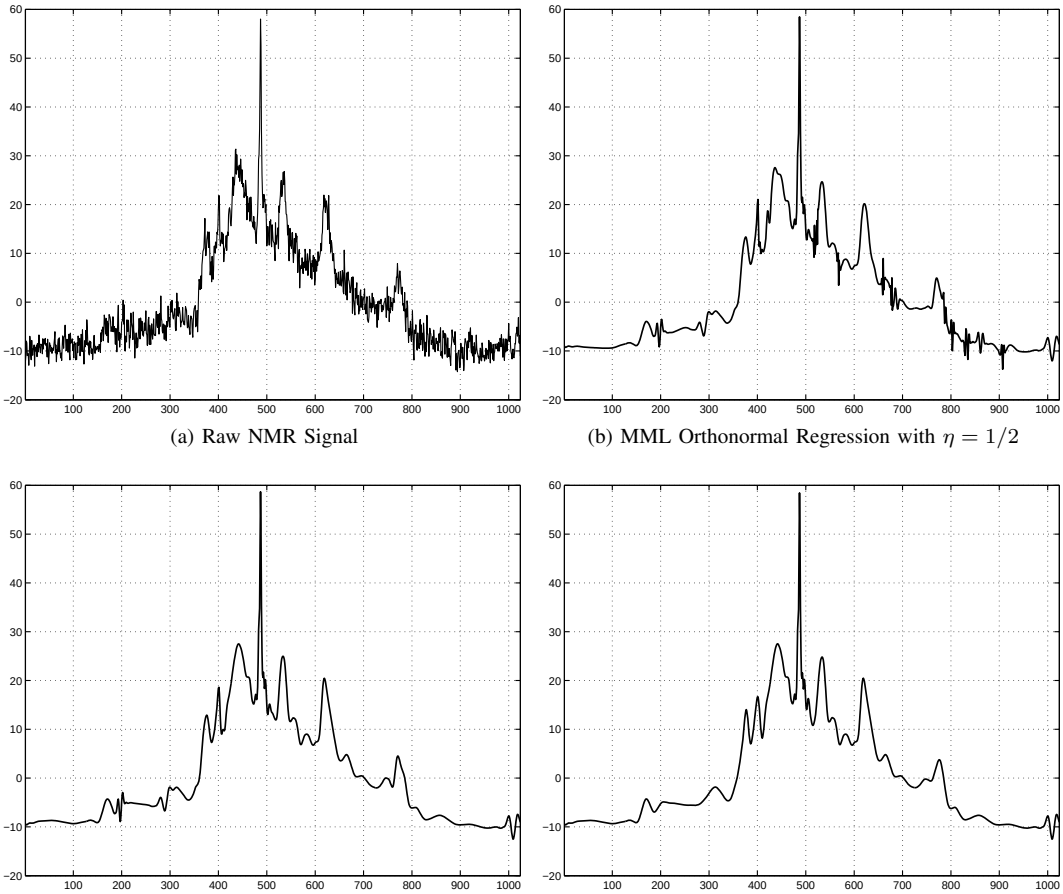
Finally, all criteria were tested on the well known NMR dataset analysed in [2]. The denoised signals for $\mathrm{MML}_{1/2}$, $\mathrm{MML}_{1/4}$ and MML_ϕ are shown in Figure 2. The MML_a and NML criteria performed similarly to $\mathrm{MML}_{1/2}$, which appears to undersmooth and allow some high frequency noise into the signal. Both the NML_r and MML_ϕ denoised signals were significantly smoother than $\mathrm{MML}_{1/2}$, with the NML_r criterion leading to a slightly smoother signal than MML_ϕ. Interestingly, in this case the $\mathrm{MML}_{1/4}$ criterion appears to have produced the best trade-off between smoothness and detail, and has included some high frequency details for $n < 512$ of the signal absent in the MML_ϕ/NML_r criteria while remaining smoother than $\mathrm{MML}_{1/2}$/NML for $n > 512$.

VII. Conclusion

This paper has derived a new criterion for orthonormal regression and wavelet denoising based on the Minimum Message Length principle. The criterion was derived within a hierarchical Bayes approach in which parameters and hyperparameters were simultaneously estimated from the data to minimise a two-part description of data and model. Under this approach parameter shrinkage was naturally incorporated into the model selection process. An asymptotic analysis of the threshold implied by this criterion showed that under certain conditions the MML procedure achieves optimal 0/1 risk. The MML orthonormal regression criterion was then applied to wavelet denoising by the adoption of suitable priors. Empirical experiments demonstrate good performance under various levels of noise. The good performance of the MML principle when applied in a hierarchical Bayes frameworks suggests a promising future application to the general linear model.

References

[1] S. G. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$," *Transactions of the American Mathematical Society*, vol. 315, no. 1, pp. 69–87, September 1989.



(a) Raw NMR Signal

(b) MML Orthonormal Regression with $\eta = 1/2$

(c) MML Orthonormal Regression with $\eta = 1/4$

(d) MML Denoising with Multilevel Prior (MML_ϕ)

Fig. 2. Denoising of NMR signal by MML

[2] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?" *Journal of the Royal Statistical Society (Series B)*, vol. 57, no. 2, pp. 301–369, 1995.

[3] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, December 1995.

[4] G. P. Nason, "Wavelet shrinkage using cross-validation," *Journal of the Royal Statistical Society (Series B)*, vol. 58, no. 2, pp. 463–479, 1996.

[5] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 173–179, March 1998.

[6] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *Journal of the Royal Statistical Society, Series B*, vol. 60, no. 4, pp. 725–749, 1998.

[7] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[8] ——, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, November 2000.

[9] ——, "Rejoinder," *The Computer Journal*, vol. 42, no. 4, pp. 343–344, 1999.

[10] E. Makalic and D. F. Schmidt, "Minimum message length shrinkage estimation," *Submitted to Statistics & Probability Letters*, 2008.

[11] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer Journal*, vol. 11, no. 2, pp. 185–194, August 1968. [Online]. Available: http://www.allisons.org/ll/MML/Structured/1968-WB-CJ/

[12] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, 1st ed., ser. Information Science and Statistics. Springer, 2005.

[13] C. Wallace and D. Boulton, "An invariant Bayes method for point estimation," *Classification Society Bulletin*, vol. 3, no. 3, pp. 11–34, 1975.

[14] P. D. Grünwald, *The Minimum Description Length Principle*, ser. Adaptive Communication and Machine Learning. The MIT Press, 2007.

[15] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *Journal of the Royal Statistical Society (Series B)*, vol. 49, no. 3, pp. 240–252, 1987.

[16] J. H. Conway and N. J. A. Sloane, *Sphere Packing, Lattices and Groups*, 3rd ed. Springer-Verlag, December 1998.

[17] C. S. Wallace and K. B. Korb, "Learning linear causal models by MML sampling," in *Causal Models and Intelligent Data Management*, A. Gammerman, Ed. Springer-Verlag, 1999, pp. 89–111.

[18] E. Makalic and D. F. Schmidt, "Efficient linear regression by minimum message length," Monash University, Tech. Rep., 2006.

[19] C. S. Wallace and P. R. Freeman, "Single-factor analysis by minimum message length estimation," *Journal of the Royal Statistical Society (Series B)*, vol. 54, no. 1, pp. 195–209, 1992.

[20] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[21] S. L. Sclove, "Improved estimators for coefficients in linear regression," *Journal of the American Statistical Association*, vol. 63, no. 322, pp. 596–606, June 1968.

[22] T. Roos, P. Myllymäki, and H. Tirri, "On the behavior of MDL denoising," in *In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005, pp. 309–316.

[23] C. M. Stein, "Confidence sets for the mean of a multivariate normal distribution," *Journal of the Royal Statistical Society (Series B)*, vol. 24, no. 2, pp. 265–296, 1962.

[24] J. B. Copas, "Regression, prediction and shrinkage," *Journal of the Royal Statistical Society (Series B)*, vol. 45, no. 3, pp. 311–354, 1983.

[25] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," September 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0609138

[26] F. Liang and A. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2708–2726, November 2004.