MDL Procedures with ℓ_1 Penalty and their Statistical Risk

Andrew R. Barron
Statistics Department
Yale University
New Haven, CT 06520-8290
Andrew.Barron@yale.edu

Xi Luo
Statistics Dept.
Yale University
New Haven, CT 06520
Xi.Luo@yale.edu

***Abstract*— We review recently developed theory for the Minimum Description Length principle, penalized likelihood and its statistical risk. An information theoretic condition on a penalty pen(f) yields the conclusion that the optimizer of the penalized log likelihood criterion log 1/likelihood(f) + pen(f) has risk not more than the index of resolvability, corresponding to the accuracy of the optimizer of the expected value of the criterion. For the linear span of a dictionary of candidate terms, we develop the validity of description-length penalties based on the ℓ_1 norm of the coefficients. New results are presented for the regression case. Other examples involve log-density estimation and Gaussian graphical statistical models.**

I. Introduction

¿From information theory and statistics the demand for high-quality data compression and accurate statistical estimation has led to the minimum description-length (MDL) principle as reviewed for instance in [7] and [12]. An application of this principle leads to a penalized likelihood criterion, optimizing $\log 1/p_f(data) + pen(f)$, where $pen(f)$ is related to the length of a description of candidate functions and $\log 1/p_f(data)$ (rounded up to an integer) is the length of the Shannon code for data given f. Building on earlier work [4], [16], in a recent paper [5] (see also [6]), joint with Cong Huang and Jonathan Li, we analyzed the statistical risk of penalized least squares and its relationship to the redundancy of data compression and exhibited the tradeoff between the accuracy of approximation and the level of complexity of candidate functions f. For this presentation (at the Workshop on Information Theoretic Methods in Science and Engineering, Tampere, Finland, August 2008) we briefly review the main conclusion of that work and give new implications for penalties based on ℓ_1 norms of coefficients in regression models based on linear combinations of terms.

The setting is as follows. The data may come from a general sample space $\underline{\mathcal{U}}$. It is traditional to think of finite length strings $\underline{U} = \underline{U}_n = (U_1, U_2, \ldots, U_n)$, consisting of a sequence of outcomes $X_1, X_2, \ldots, X_n$ or outcome pairs $(X_i, Y_i)_{i=1}^n$. We write $P_{\underline{U}|f}$ (or more briefly P_f) for the distributions on $\underline{U}$ indexed by functions f in some collection $\mathcal{F}$. Likewise $E_{\underline{U}|f}$ or more briefly E_f denotes the expected value. When being explicit about sample size, we index by n, as in $P_{\underline{U}_n|f}$ or $P_f^{(n)}$. These distributions are assumed to have density functions $p(\underline{u}|f) = p_f(\underline{u})$, relative to a fixed reference measure, which provides the likelihood function of f at data $\underline{U}$. The reference measure is assumed to be a product of measures on the individual spaces. For the special case of i.i.d. modeling, there is a space $\mathcal{U}$ for the individual outcomes with distributions $P_f^{(1)} = P_f$ and then $\underline{\mathcal{U}}$ is taken to be the product space $\mathcal{U}^n$ and $P_{\underline{U}_n|f} = P_f^n$ is taken to be the product measure with joint density $p_f(\underline{u}_n) = \prod_{i=1}^n p_f(u_i)$.

The Kullback divergence and a Bhattacharyya, Rényi, Hellinger divergence are used in examining the quality of statistical estimates and data compression. The Kullback divergence $D(P_{\underline{U}}\|Q_{\underline{U}}) = E \log p(\underline{U})/q(\underline{U})$ is the total expected redundancy for data $\underline{U}$ described using $q(\underline{u})$ but governed by a density $p(\underline{u})$. Likewise the Bhattacharyya, Hellinger, Rényi divergence [8], [11], [19] is given by $d(P_{\underline{U}}, Q_{\underline{U}}) = 2 \log 1/\int (p(\underline{u})q(\underline{u}))^{1/2}$. We use $D_n(f^*, f)$ and $d_n(f^*, f)$ to denote the divergences between the joint distributions $P_{\underline{U}|f^*}$ and $P_{\underline{U}|f}$. In the i.i.d. modeling case these take the form $D_n(f^*, f) = nD(f^*, f)$ and $d_n(f^*, f) = nd(f^*, f)$, respectively, where $D(f^*, f)$ and $d(f^*, f)$ are the divergences between the single observation distributions $P_{U_1|f^*}$ and $P_{U_1|f}$. The divergences measure how well f approximates f^*.

Writing $D(P||Q) = -2E \log(q(U)/p(U))^{1/2}$ and employing Jensen's inequality shows that $D(P\|Q) \geq d(P, Q)$. The relationship to the squared Hellinger distance $H^2(P, Q) = \int (p(u)^{1/2} - q(u)^{1/2})^2$ is $d(P, Q) = -2 \log(1 - \frac{1}{2}H^2)$ which is not less than $H^2(P, Q)$. These divergences upper bound the square of the L_1 distance. Moreover, $d(P, Q)$ is locally equivalent to the Kullback-Leibler divergence when $\log p(u)/q(u)$ is upper-bounded by a constant. Moreover, it evaluates to familiar quantities in special cases, e.g., for two normals of mean μ and $\tilde{\mu}$ and variance σ^2, it is $\frac{1}{4}(\mu - \tilde{\mu})^2/\sigma^2$. The most important reason for our use of the Bhattacharyya, Rényi, Hellinger loss function is that it allows clean examination of the risk, without putting any conditions on the density functions $p_f(\underline{u})$.

Two-stage codes were used in the original formulation of the MDL principle [20], [21] and in the analysis of [4]. One works with a countable set of possible functions, perhaps obtained by discretization of the underlying family $\mathcal{F}$. In this case the requirement on the penalty is that it corresponds to the length of a description, which means satisfaction of the Kraft inequality $\sum_f 2^{-pen(f)} \leq 1$. Then, for each function f and data $\underline{U}$, one has a two-stage codelength $pen(f) + \log 1/p_f(\underline{U})$

corresponding to the bits of description of f followed by the bits of the Shannon code for $\underline{U}$ given f. Then the minimum total two-stage codelength takes the form

$$\min_f \left\{ \log \frac{1}{p_f(\underline{U})} + pen(f) \right\}.$$

A minimizer $\hat{f}$ is called the minimum complexity estimator for density estimation [4] and it is also called the complexity regularization estimator for regression and classification problems [1].

Typical behavior of the minimal two stage codelength is revealed by investigating what happens when the data $\underline{U}_n$ are distributed according to $p_{f^*}(\underline{u}_n)$ for various possible f^*. It is helpful to have a notion of a surrogate function f_n^*, appropriate to the current sample size n, which best resolves f^*. The appropriateness of such an f_n^* is judged by whether it captures expected compression and estimation properties of the target.

The redundancy rate of the two-stage description is shown in [4] to be not more than the index of resolvability defined by

$$R_n(f^*) \, = \, \min_f \left\{ \frac{1}{n} D(P_{\underline{U}_n|f^*} || P_{\underline{U}_n|f}) + \frac{1}{n} pen(f) \right\}.$$

For i.i.d. modeling it takes the form

$$R_n(f^*) \, = \, \min_f \left\{ D(f^*, f) + \frac{pen(f)}{n} \right\},$$

capturing the ideal tradeoff in error of approximation of f^* and the complexity relative to the sample size. The function f_n^* which achieves this minimum is the population counterpart to the sample-based $\hat{f}$. It best resolves the target for the given sample size. Since $\hat{f}$ is the sample-based minimizer, one has an inequality between the pointwise redundancy and a pointwise version of the resolvability

$$\log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + L_n(\hat{f}) \;\; \leq \;\; \log \frac{p_{f^*}(\underline{U})}{p_{f_n^*}(\underline{U})} + L_n(f_n^*).$$

The resolvability bound on the expected redundancy is the result of taking the expectation of this pointwise inequality.

This $R_n(f^*)$ also bounds the statistical risk of $\hat{f}$, as we recall. First we recall such a risk bound for penalized likelihood with a countable set $\tilde{\mathcal{F}}$ of candidate functions. Henceforth we use base e exponentials and logarithms to simplify the mathematics (the units for coding interpretations become nats rather than bits). The following result originates in Jonathan Li's thesis [16] and is proven also in [5], [6], [15], [12].

Theorem 1.1: Resolvability bound on risk. For a countable $\tilde{\mathcal{F}}$, suppose $pen(f) \, \geq \, 2L_n(f)$ where $L_n(f)$ satisfies $\sum_{f \in \tilde{\mathcal{F}}} e^{-L_n(f)} \leq 1$ and let $\hat{f}$ be the estimator achieving

$$\min_{f \in \tilde{\mathcal{F}}} \left\{ \log \frac{1}{p_f(\underline{U}_n)} \, + \, pen(f) \right\}.$$

Then, for any target function f^* and for all sample sizes, the expected divergence of $\hat{f}$ from f^* is bounded by the index of resolvability

$$Ed_n(f^*, \hat{f}) \; \leq \; \min_{f \in \tilde{\mathcal{F}}} \{ \, D_n(f^*, f) + pen(f) \, \}.$$

In particular with i.i.d. modeling, the risk satisfies

$$Ed(f^*, \hat{f}) \; \leq \; \min_{f \in \tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{pen(f)}{n} \right\}.$$

Corollary 1.2: If, in the i.i.d. case, the log density ratios are bounded by a constant B, that is, if $|\log p_{f^*}(u)/p_f(u)| \leq B$ for all $f \in \tilde{\mathcal{F}}$, then there is a constant $C_B \leq 2 + B$ such that the Kullback risk satisfies

$$ED(f^*, \hat{f}) \leq C_B \min_{f \in \tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{pen(f)}{n} \right\}.$$

Corresponding results for uncountable $\mathcal{F}$ were developed in [5], [6], allowing application to optimization over real-valued parameters in standard statistical models. In that analysis an important role is played by a measure of the discrepancy between empirical and population values of the log-likelihood ratio at a candidate f. As explained there it is given by

$$dis(f) = \log \frac{p_{f^*}(\underline{U})}{p_f(\underline{U})} - 2 \log \frac{1}{E(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2}}.$$

In the proof of Theorem 1.1 from the countable case, if an information-theoretically valid penalty $pen(f)$ is added to the discrepancy, then uniformly in f (i.e., even with a data-based $\hat{f}$ in place of a fixed f) the expectation of the penalized discrepancy is positive.

This leads to consideration, in the uncountable case, of penalties which exhibit a similar discrepancy control. We say that a collection $\mathcal{F}$ with a penalty $pen(f)$ for $f \in \mathcal{F}$ has a *variable–complexity variable–discrepancy cover* suitable for p_{f^*} if there exists a countable $\tilde{\mathcal{F}}$ and $\mathcal{L}(\tilde{f}) = 2L(\tilde{f})$ satisfying $\sum_{\tilde{f}} e^{-L(\tilde{f})} \leq 1$, such that the following condition $(*)$ holds for all $\underline{U}$:

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ \, dis(\tilde{f}) \; + \; \mathcal{L}(\tilde{f}) \, \right\} \leq \inf_{f \in \mathcal{F}} \left\{ \, dis(f) \; + \; pen(f) \, \right\}. \tag{$*$}$$

This condition captures the aim that the penalty in the uncountable case mirrors an information-theoretically valid penalty in the countable case. The above condition gives what we want because the minimum over the countable $\tilde{f}$ is shown to have non-negative expectation and so the minimum over all f in $\mathcal{F}$ will also.

Equivalent to condition $(*)$ the following characterization $(**)$ is convenient. For each f in $\mathcal{F}$ there is an associated representer $\tilde{f}$ in $\tilde{\mathcal{F}}$ for which

$$pen(f) \; \geq \; \log \frac{p_f(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2 \log \frac{E(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2}}{E(p_{\tilde{f}}(\underline{U})/p_{f^*}(\underline{U}))^{1/2}} + 2L(\tilde{f}) \, . \tag{$**$}$$

The idea is that if $\tilde{f}$ is close to f then the discrepancy difference is small. Then the complexity of such $\tilde{f}$ along with the discrepancy difference assesses whether a penalty $pen(f)$ is suitable. The minimizer in $\tilde{\mathcal{F}}$ depends on the data and accordingly we allow the representer $\tilde{f}$ of f to also have such dependence. With this freedom, in cases of interest, the variable complexity cover condition indeed holds for all $\underline{U}$, though it would suffice for our purposes that (*) hold in expectation.

One strategy to verify the condition would be to create a metric-based cover of $\mathcal{F}$ with a metric chosen such that for each f and its representer $\tilde{f}$ one has $|\log p_f(\underline{U})/p_{\tilde{f}}(\underline{U})|$ plus the difference in the divergences arranged if possible to be less than a distance between f and $\tilde{f}$. Some examples where this can be done are in [4]. Such covers give a metric entropy flavor, though the $L(\tilde{f})$ provides variable complexity rather than the fixed log-cardinality of metric entropy.

Condition $(**)$ specifies that there be a cover with variable distortion plus complexity rather than a fixed distance and fixed cardinality. This is analogous to the distortion plus rate tradeoff in Shannon's rate-distortion theory. In our treatment, the distortion is the discrepancy difference (which does not need to be a metric), the codebook is the cover $\tilde{\mathcal{F}}$, the codelengths are the complexities $L(\tilde{f})$. Valid penalties $pen(f)$ exceed the minimal sum of distortion plus complexity.

The generalization of Theorem 1.1 to the case of uncountable $\mathcal{F}$, is the following.

Theorem 1.3: Consider $\mathcal{F}$ and $pen(f)$ satisfying the discrepancy plus complexity requirement $(*)$ and the estimator $\hat{f}$ achieving the optimum penalized likelihood

$$\min_{f\in\mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{U})} + pen(f) \right\}.$$

If the data $\underline{U}$ are distributed according to $P_{\underline{U}|f^*}$, then

$$Ed_n(f^*,\hat{f}) \;\le\; \min_{f\in\mathcal{F}} \left\{ E\log \frac{p_{f^*}(\underline{U})}{p_f(\underline{U})} \;+\; pen(f) \right\}.$$

In particular, for i.i.d. modeling,

$$Ed(f^*,\hat{f}) \;\le\; \min_{f\in\mathcal{F}} \left\{ D(f^*,f) + \frac{pen(f)}{n} \right\}.$$

II. Information-theoretic validity of ℓ_1 penalty for log-densities

Before giving the new implication for regression, we recall first in this section the implication for log-density estimation.

In this case f models the log density function of independent random variables $X_1,\ldots,X_n$, in the sense that for some reference density $p_0(x)$ we have

$$p_f(x) = \frac{p_0(x)\, e^{\,f(x)}}{c_f}$$

where c_f is the normalizing constant. Examining the difference in discrepancies at f and a representing $\tilde{f}$ we see that both $p_0(x)$ and c_f cancel out. What remains for our penalty requirement is that for each f in $\mathcal{F}$ there is a $\tilde{f}$ in a countable $\tilde{\mathcal{F}}$ with complexities $L(\tilde{f})$ for which

$$pen(f) \;\ge\; 2L(\tilde{f}) + \sum_{i=1}^{n} (f(X_i) - \tilde{f}(X_i)) + 2n \log E \exp\{\tfrac{1}{2}(\tilde{f}(X) - f(X))\}$$

where the expectation is with respect to a distribution for X constructed to have density which is the normalized pointwise affinity $p_a(x) = [p_{f^*}(x)p_f(x)]^{1/2}/A(f^*,f)$.

In this section we illustrate how to demonstrate the existence of such representers $\tilde{f}$ using an ℓ_1 penalty on coefficients in representation of f in the linear span of a dictionary of candidate basis functions.

Let $\mathcal{F}$ be the linear span of a dictionary $\mathcal{H}$ of functions. Thus any f in $\mathcal{F}$ is of the form $f(x) = f_\theta(x) = \sum_h \theta_h h(x)$ where the coefficients are denoted $\theta = (\theta_h : h \in \mathcal{H})$. We assume that the functions in the dictionary are bounded. We want to show that a weighted ℓ_1 norm of the coefficients $||\theta||_1 = \sum_h |\theta_h| a_h$ can be used to formulate a valid penalty. Here we use the weights $a_h = \|h\|_\infty$. For f in $\mathcal{F}$ we denote $V_f = \min\{\|\theta\|_1 : f_\theta = f\}$. With the definition of V_f further extended to a closure of $\mathcal{F}$, this V_f is called the variation of f with respect to $\mathcal{H}$. We will show that certain multiples of V_f are valid penalties.

The dictionary $\mathcal{H}$ is a finite set of p candidate terms, typically much larger than the sample size. As we shall see, the codelengths of our representers will arise via a variable number of terms times the log cardinality of the dictionary. Accordingly, for sensible risk bounds, it is only the logarithm of p, and not p itself, that we need to be small compared to the sample size n.

A valid penalty will be seen to be a multiple of V_f, by arranging the number of terms in the representer to be proportional to V_f and by showing that a representer with that many terms suitably controls the discrepancy difference. We proceed now to give the specifics.

The countable set $\tilde{\mathcal{F}}$ of representers is taken to be the set of all functions of the form $\tilde{f}(x) = V\frac{1}{K}\sum_{k=1}^{K} h_k(x)/a_{h_k}$ for terms h_k in $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, where the number of terms K is in $\{1,2,\ldots\}$ and the nonnegative multipliers V will be determined from K in a manner we will specify later. We let p be the cardinality of $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, allowing for h or $-h$ or 0 to be a term in $\tilde{f}$ for each h in $\mathcal{H}$.

The main part of the codelength $L(\tilde{f})$ is $K \log p$ nats to describe the choices of $h_1,\ldots,h_K$. The other part is for the description of K and it is negligible in comparison, but to include it simply, we may use a possibly crude codelength for the integer K such as $K \log 2$. Adding these contributions of $K \log 2$ for the description of K and of $K \log p$ for the description of $\tilde{f}$ given K, we have

$$L(\tilde{f}) = K \log(2p).$$

To establish existence of a representer $\tilde{f}$ of f with the desired properties, we put a distribution on choices of $h_1,\ldots,h_K$ in which each is selected independently, where h_k is h with probability $|\theta_h|a_h/V$ (with a sign flip if θ_h is negative). Here $K = K_f = \lceil V_f/\delta \rceil$ is set to equal V_f/δ rounded up to the nearest integer, where $V_f = \sum_h |\theta_h| a_h$, where a small value for δ will be specified later. Moreover, we set $V = K\delta$, which is V_f rounded up to the nearest point in a grid of spacings δ. When V_f is strictly less than V there is leftover an event of probability $1 - V_f/V$ in which h_k is set to 0.

As f varies, so does the complexity of its representers. Yet for any one f, with $K = K_f$, each of the possibilities for

the terms h_k produces a possible representer $\tilde{f}$ with the same complexity $K_f \log 2p$.

The key property of our random choice of $\tilde{f}(x)$ representing $f(x)$ is that, for each x, it is a sample average of i.i.d. choices $Vh_k(x)/a_{h_k}$. Each of these terms has expectation $f(x)$ and variance $V\sum_h |\theta_h| h^2(x)/a_h - f^2(x)$ not more than V^2.

As the sample average of K such independent terms, $\tilde{f}(x)$ has expectation $f(x)$ and variance $(1/K)$ times the variance given for a single draw. We will also need expectations of exponentials of $\tilde{f}(x)$ which is made possible by the representation of such an exponential of sums as the product of the exponentials of the independent summands.

The existence argument proceeds as follows. The quantity we need to bound to set a valid penalty is the minimum over $\tilde{\mathcal{F}}$ of the complexity-penalized discrepancy difference:

$$2L(\tilde{f}) + \sum_{i=1}^{n}(f(X_i) - \tilde{f}(X_i)) + 2n \log \int p(x) e^{(\tilde{f}(x) - f(x))/2}$$

where $p(x) = p_a(x)$ is a probability density function as specified in the preceding section. The minimizing $\tilde{f}$ gives a value not more than the expectation over random $\tilde{f}$ obtained by the sample average of randomly selected h_k. We condition on the data $X_1, \ldots X_n$. The terms $f(X_i) - \tilde{f}(X_i)$ have expectation 0 so it remains to bound the expectation of the log term. It is less than or equal to the log of the expectation, so we bring that expectation inside the integral. Then at each x we examine the expectation of the exponential of $\frac{1}{2}[\tilde{f}(x) - f(x)]$. By the independence and identical distribution of the K summands that comprise the exponent, the expectation is equal to the Kth power of the expectation of $\exp\{\frac{1}{2K}[Vh(x)/a_h - f(x)]\}$ for a randomly drawn h.

We now take advantage of classical bound of Hoeffding, easily verified by using the series expansion of the exponential. If T is a random variable with range bounded by B, then $E\exp\{\frac{1}{K}(T - \mu)\} \leq \exp\{\frac{B^2}{8K^2}\}$.

Let $R(x) = \max_h h(x)/a_h - \min_h h(x)/a_h$ be the range of $h(x)/a_h$ as h varies for the given x, which is uniformly bounded by 2. At x given, $T = \frac{1}{2}Vh(x)/a_h$ is a random variable, induced by the random h, having range $\frac{V}{2}R(x)$. Then at the given x, using the Hoeffding inequality gives that the expectation of $\exp\{\frac{1}{2}(\tilde{f}(x) - f(x))\}$ is bounded by $\exp\{\frac{(VR(x))^2}{32K}\}$ which is not more than $\exp\{\frac{V^2}{8K}\}$.

The expectation of the log of the integral of this exponential is bounded by $\frac{V^2}{8K}$ or equivalently $\frac{1}{8}V\,\delta$. When multiplied by $2n$, it yields a discrepancy difference bound of

$$\tfrac{1}{4} n\, V\, \delta,$$

where V is not more than $V_f + \delta$.

Now twice the complexity plus the discrepancy bound has size $2K\log(2p) + \frac{1}{4}nV_f\delta + \frac{1}{4}n\delta^2$, which, with our choice of $K = \lceil V_f/\delta \rceil$ not more than $V_f/\delta + 1$, shows that a penalty of the form

$$pen_n(f) \;\geq\; \lambda V_f \;+\; C$$

is valid as long as λ is at least $\frac{2}{\delta}\log(2p) + \frac{1}{4}n\delta$ and $C = 2\log(2p) + \frac{1}{4}n\delta^2$. We set $\delta = (\frac{8\log 2p}{n})^{1/2}$ as it optimizes the bound on λ producing a critical value λ_n^* equal to $(2n\log 2p)^{1/2}$ and a value of $C = 4\log(2p)$. The presence of the constant term C in the penalty does not affect the optimization that produces the penalized likelihood estimator, that is, the estimator is the same as if we used a pure ℓ_1 penalty equal to λV_f. Nevertheless, for application of our theory giving risk bounds, the C found here is part of our bound.

We summarize the conclusion with the following Theorem. The setting is as above with the density model $p_f(x)$ with exponent $f(x)$. The estimate is chosen with f in the linear span of the dictionary $\mathcal{H}$. The data are i.i.d. according to $p_{f^*}(x)$.

Theorem 2.1: The ℓ_1 penalized likelihood estimator $\hat{f} = f_{\hat{\theta}}$ achieving

$$\min_{\theta} \left\{ \log \frac{1}{p_{f_\theta}(\underline{X}_n)} + \lambda_n ||\theta||_1 \right\},$$

or, equivalently,

$$\min_{f} \left\{ \log \frac{1}{p_{f}(\underline{X}_n)} + \lambda_n V_f \right\},$$

has risk $Ed(f^*, \hat{f})$ bounded for every sample size by

$$R_n(f^*) \;\leq\; \inf_{f \in \mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} \right\} \;+\; \frac{4\log 2p}{n}$$

provided $\frac{\lambda_n}{n} \geq \left[\frac{2\log(2p)}{n}\right]^{1/2}$.

In particular, if f^* has finite variation V_{f^*} then for all n,

$$Ed(f^*, \hat{f}) \;\leq\; R_n(f^*) \leq \frac{\lambda_n V_{f^*}}{n} + \frac{4\log 2p}{n}.$$

Note that the last term $\frac{4\log 2p}{n}$, is typically negligible compared the main term, which is near

$$\left[\frac{2\log 2p}{n}\right]^{1/2} V_{f^*}\,.$$

Not only does this result exhibit $[(\log p)/n]^{1/2}$ as the rate of convergence, but also it gives clean finite sample bounds.

Even if V_{f^*} is finite, the best resolvability can occur with simpler functions. In fact, until n is large compared to $V_{f^*}^2 \log p$, the index of resolvability will favor approximating functions f_n^* with smaller variation.

In this section we have demonstrated the validity of an ℓ_1 penalty for log-densities for bounded functions in the dictionary. We would also like to deal with unbounded functions satisfying certain moment conditions, as arises in the Gaussian graphical models. In a separate work by Xi Luo, ℓ_1 is also a valid penalty for such Gaussian graphical models verified using Bernstein's moment condition valid for the Gaussian distribution.

III. Information-theoretic validity of ℓ_1 penalty for regression

Now consider the linear regression case with fixed design. At each x_i we seek a fit $f(x_i)$ to a corresponding outcome

Y_i. We use the Gaussian model of independent outcome $Y_1, \ldots, Y_n$ with joint density function

$$p_f(\underline{y}|\underline{x}) = \frac{1}{(2\pi\sigma^2)^{(n/2)}} exp\left\{ -\frac{\sum_{i=1}^n (y_i - f(x_i))^2}{2\sigma^2} \right\}.$$

The case of fixed (known) variance σ^2 is considered first. In this setting, the divergence $d(P_{\underline{Y}|\underline{x},f^*}, P_{\underline{Y}|\underline{x},f})$ for fixed $\underline{x}$ can be written explicitly as

$$\frac{1}{4\sigma^2}\sum_{i=1}^n (f(x_i) - f^*(x_i))^2.$$

Then in accordance with $(**)$ we check validity of a penalty $pen(f)$ by verifying for a suitable representer $\tilde{f}$ that

$$pen(f) \geq 2L(\tilde{f}) + \frac{1}{2\sigma^2}\sum_{i=1}^n \left[(y_i - \tilde{f}(x_i))^2 - (y_i - f(x_i))^2 \right]$$
$$- \frac{1}{4\sigma^2}\sum_{i=1}^n \left[(f^*(x_i) - \tilde{f}(x_i))^2 - (f^*(x_i) - f(x_i))^2 \right].$$

In this section we adapt the general strategy developed in the previous section to the regression setting to demonstrate that the ℓ_1 penalty on coefficients with suitable multipliers is also an information-theoretic penalty for regression. The result presented here is fascinating for us as it also reveals what penalty parameter λ should employed for ℓ_1 penalized regression to be justifiable for the MDL interpretation and statistical risk analysis.

We allow the weights a_h in this section to be empirical ℓ_2 norm $\|h\|_{\underline{x}}$ where $\|h\|_{\underline{x}}^2 = \frac{1}{n}\sum_i h(x_i)$ instead of $\|h\|_\infty$ in the previous section. we no longer need a bounded range condition nor an appeal to the Hoeffding inequality. The same sampling strategy for generating a random $\tilde{f}$ also applies here.

We bound similarly the minimum over $\tilde{\mathcal{F}}$ of the complexity-penalized discrepancy difference by the quantity obtained by the sample average of randomly selected h_k. For the discrepancy difference, adding and subtracting $f(x_i)$ in each square, the squared terms of $y_i - f(x_i)$ and $f^*(x_i) - f(x_i)$ cancel out when expanding out the squares and their cross product terms with $(f(x_i) - \tilde{f}(x_i))$ vanish in expectation under the random $\tilde{f}(x_i)$. What remains for the expected discrepancy difference is the expectation of

$$\frac{1}{4\sigma^2}\sum_{i=1}^n (\tilde{f}(x_i) - f(x_i))^2.$$

Each summand $(\tilde{f}(x_i) - f(x_i))^2$ for fixed x_i under random $\tilde{f}$ has mean not more than the $(1/K)$ times the bound $V\sum_h |\theta_h| h^2(x_i)/a_h$ on the variance given for a single draw h. The aggregated bound over x_i yields

$$\frac{V}{4\sigma^2 K}\sum_{i=1}^n \sum_h |\theta_h| h^2(x_i)/a_h = \frac{nVV_f}{4\sigma^2 K}$$

where n appearing in the equality is by the fact that $\sum_{i=1}^n h^2(x_i)/a_h^2 = n$ for each h.

Now the discrepancy difference plus twice the complexity penalty is bounded by

$$2K\log(2p) + \frac{nVV_f}{4\sigma^2 K}.$$

With our choice of $K = \lceil V_f/\delta \rceil = V/\delta$ not more than $V_f/\delta + 1$, we show that the penalty of the form

$$pen(f) \geq \lambda V_f + C$$

is valid as long as λ is not smaller than $2V_f(\log 2p)/\delta + nV_f\delta/(4\sigma^2)$ and $C = 2\log(2p)$. Setting $\delta = 2\sigma(2(\log 2p)/n)^{(1/2)}$ to optimize the bound for λ, the critical value is $\lambda^* = (2n\log(2p))^{(1/2)}/\sigma$ and our analysis shows that ℓ_1 is valid as long as the penalty parameter exceeds λ^*.

Consequently we have a simple risk bound for this regression setting with fixed design and known variance σ^2. In particular, the Kullback divergence and Bhattacharyya, Rényi, Hellinger divergence measuring the density distance can be explicitly written in the form of squared errors.

Theorem 3.1: The ℓ_1 penalized least squares estimator $\hat{f} = f_{\hat{\theta}}$ achieving

$$\min_\theta \left\{ \frac{1}{n}\sum_{i=1}^n (y_i - f_\theta(x_i))^2 + 2\sigma\frac{\lambda_n}{n}||\theta||_1 \right\}$$

has the following risk bound

$$E\|\hat{f} - f^*\|_{\underline{x}}^2$$
$$\leq 2\inf_\theta \left\{ \|f_\theta - f^*\|_{\underline{x}}^2 + 2\sigma\frac{\lambda_n}{n}\|\theta\|_1 \right\} + \frac{8\sigma^2\log(2p)}{n}$$

provided that $\frac{\lambda_n}{n} \geq \left[\frac{2\log(2p)}{n}\right]^{1/2}$.

Next we generalize the result to the unknown σ case. Following the MDL principle we are motivated to estimate $(\hat{f}, \hat{\sigma}^2)$ by optimizing

$$\frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \frac{1}{2}\log 2\pi\sigma^2$$
$$+ \frac{1}{\sigma}\frac{\lambda_n}{n}\|\theta\|_1 + \frac{pen(\sigma^2)}{n}$$

where the first two terms form the $-\frac{1}{n}\log likelihood$ and the next term is the penalty used in the fixed σ^2 case. With $pen(f, \sigma^2)$ similar to this, we show that such an optimization indeed satisfies the requirement $(**)$ for validity of statistical risk analysis. For the representer $(\tilde{f}, \tilde{\sigma}^2)$ in a countable cover, we adapt the same strategy of random K-term $\tilde{f}$ and use for $\tilde{\sigma}^2$ of a logarithmic discretization of σ^2, that is, $\log\tilde{\sigma}^2 = \lfloor(\log\sigma^2)/\epsilon\rfloor\epsilon = K'\epsilon$ with the choice of ϵ to be specified and K' an integer. We set the codelength in this case to be $L(\tilde{f}, \tilde{\sigma}^2) = K\log(2p) + 2\log(K'+1)$ where we crudely encode K' by $2\log(K'+1)$ for simplicity. The Bhattacharyya, Rényi, Hellinger divergence $d(P_{\underline{Y}|\underline{X},f^*,\sigma_*}, P_{\underline{Y}|\underline{X},f,\sigma})$ can be written explicitly as

$$\frac{1}{2(\sigma^2 + {\sigma_*}^2)}\sum_{i=1}^n (f(x_i) - f^*(x_i))^2 + \log\frac{\sigma^2 + \sigma_*^2}{2\sqrt{\sigma^2\sigma_*^2}}.$$

Now checking $(**)$ involves the difference of these divergences at (f,σ^2) and at $(\tilde f,\tilde\sigma^2)$ as well as the differences in the log-likelihood. Some of the resulting terms in the difference are negative (can can be dropped) by our choice of $\tilde\sigma^2$ not more than σ^2 (by rounding down). What remains to verify is that

$$pen(f,\sigma^2) \geq 2L(\tilde f,\tilde\sigma^2) + \sum_{i=1}^n \left[\frac{(y_i-\tilde f(x_i))^2}{2\tilde\sigma^2} - \frac{(y_i-f(x_i))^2}{2\sigma^2}\right] - \sum_{i=1}^n \left[\frac{(f^*(x_i)-\tilde f(x_i))^2}{2(\sigma_*^2+\tilde\sigma^2)} - \frac{(f^*(x_i)-f(x_i))^2}{2(\sigma_*^2+\sigma^2)}\right].$$

To show existence of a suitable representer $\tilde f$ we bound again the sample average version. The same bound for $(\tilde f(x_i)-f(x_i))^2$ is used and we drop all non-positive terms for cleanness. The discrepancy difference plus twice the complexity is then bounded by

$$\frac{1}{2\sigma^2}\left[(e^\epsilon-1)\sum_{i=1}^n (y_i-f(x_i))^2 + e^\epsilon \frac{nVV_f}{K}\right] + 2K\log(2p) + 2\log(K'+1).$$

With $K' = \lfloor(\log\sigma^2)/\epsilon\rfloor \leq (\log\sigma^2)/\epsilon$ and $K = V/\delta \leq V_f/\delta + 1$, we set $\delta = 2\sigma((\log 2p)/n)^{1/2}e^{-\epsilon/2}$ to optimize the bound assuming ϵ fixed first. For simplicity, we pick $\epsilon = 1/(2n)$ to optimize over ϵ crudely and use $e^{1/2n} < 1+1/n$ to simplify the multiplying constants. The resulting satisfactory penalty requirement takes the form

$$pen(f,\sigma^2) \geq \frac{1}{2\sigma^2}\|y-f\|_{\underline x}^2 + \frac{\lambda V_f}{\sigma} + 2\log\sigma^2 + 2\log(4pn),$$

valid as long as $\lambda \geq (2+\frac{1}{n})\sqrt{n\log(2p)}$, where we denote $\|y-f\|_{\underline x}^2 = \frac{1}{n}\sum_{i=1}^n (y_i-f(x_i))^2$. The main part of this penalty is the $\lambda V_f/\sigma$ term with the $\sqrt{n}$ factor; the other terms are of lower order. Recall that for $f=f_\theta$ the V_f is determined by the ℓ_1 norm $\|\theta\|_1$.

Consequently we have the following theorem.

Theorem 3.2: The ℓ_1 penalized least squares estimator $\hat f = f_{\hat\theta}$ achieving

$$\min_{\theta,\sigma}\left\{\frac{1}{2\sigma^2}(1+\frac{1}{n})\frac{1}{n}\sum_{i=1}^n (y_i-f_\theta(x_i))^2 + \frac{\lambda_n}{n\sigma}||\theta||_1 + \frac{\log(\sigma^2)}{2}(1+\frac{4}{n})\right\},$$

has the following risk bound

$$\frac{1}{n}Ed(P_{\underline{Y}|\underline x,f^*,\sigma_*}, P_{\underline{Y}|\underline x,\hat f,\hat\sigma}) \leq \inf_{\theta,\sigma^2}\left\{\frac{1}{n}D(P_{\underline{Y}|\underline x,f^*,\sigma_*}||P_{\underline{Y}|\underline x,f_\theta,\sigma}) + \frac{\lambda_n}{n\sigma}\|\theta\|_1 + \frac{\|y-f_\theta\|_{\underline x}^2}{2\sigma^2 n} + \frac{2\log\sigma^2}{n}\right\} + \frac{2\log(4pn)}{n}.$$

provided that $\frac{\lambda_n}{n} \geq (2+\frac{1}{n})\sqrt{\log(2p)/n}$.

COMMENT ON COMPUTATION FOR REGRESSION: The optimization producing $\hat\theta,\hat\sigma^2$ in Theorem 3.2 is reasonably straightforward. Each value of σ^2 corresponds to a multiplier of the ℓ_1 penalty. For each value of σ^2 one may optimize over θ by standard ℓ_1-penalized least squares algorithms. A particularly fast such method with computational guarantees is the greedy algorithm in the manuscript [13], the log-likelihood version of it is in section IV below. Then rather than picking the multiplier by some auxiliary cross-validation method, MDL chooses it (or equivalently chooses the single parameter σ^2) to optimize the above criterion.

Alternatively, we note that for θ the best $\sigma^2 = \sigma_{f_\theta}^2$ solves a quadratic

$$(1+\frac{4}{n})\sigma^2 = \sigma\frac{\lambda_n}{n}\|\theta\|_1 + (1+\frac{1}{n})\|y-f_\theta\|_{\underline x}^2.$$

Whence one may plug in the solution $\sigma_{f_\theta}^2$ and optimize the resulting function of θ above.

IV. GREEDY COMPUTATION FOR LOG-DENSITIES ESTIMATION

Again for log-density estimation, consider a relaxed greedy algorithm in which we successively optimize the ℓ_1 penalized likelihood one term at a time, optimizing choices of α, β and h in the update

$$\hat f_k(x) = (1-\alpha)\hat f_{k-1}(x) + \beta h(x)$$

for each $k = 1,2,\ldots$. Our result is that it solves the ℓ_1 penalized likelihood optimization, with a guarantee that after k steps we have a k component mixture within order $1/k$ of the optimum. Similar results for ℓ_1 penalized least squares are in [13]. Indeed, one initializes with $\hat f_0(x) = 0$ and $v_0 = 0$. Then for each step k, ones optimizes α, β, and h to provide the the kth term $h_k(x)$. At each iteration one loops through the dictionary trying each $h\in\mathcal H$, solving for the best associated scalars $0\leq\alpha\leq 1$ and $\beta\in R$, and picks the h that best improves the ℓ_1 penalized log-likelihood, using $v_k = (1-\alpha)v_{k-1} + |\beta|\ a_{h_k}$ as the updated bound on the variation of $\hat f_k$. This is a case of what we call an ℓ_1 *penalized greedy pursuit*. This algorithm solves the penalized log-likelihood problem, with an explicit guarantee on how close we are to the optimum after k steps. Indeed, for any given data set $\underline X$ and for all $k\geq 1$,

$$\frac{1}{n}\left[\log\frac{1}{p_{\hat f_k}(\underline X)} + \lambda v_k\right] \leq \inf_f\left\{\frac{1}{n}\left[\log\frac{1}{p_f(\underline X)} + \lambda V_f\right] + \frac{2V_f^2}{k+1}\right\},$$

where the infimum is over functions in the linear span of the dictionary, and the variation corresponds to the weighted ℓ_1 norm $\|\theta\|_1 = \sum_{h\in\mathcal H}|\theta_h|a_h$, with a_h set to be not less than $\|h\|_\infty$. This inequality shows that $\hat f_k$ has penalized log-likelihood within order $1/k$ of the optimum.

This computation bound for ℓ_1 penalized log-likelihood is developed in the Yale thesis research of one of us, Xi Luo, adapting some ideas from the corresponding algorithmic theory for ℓ_1 penalized least squares from [13]. The proof of this computation bound and the risk analysis given above have aspects in common. So it is insightful to give the proof here.

It is equivalent to show that for each f in the linear span that

$$\frac{1}{n}\left[\log\frac{p_f(\underline{X}_n)}{p_{\hat{f}_k}(\underline{X}_n)}+\lambda(v_k-V_f)\right] \;\le\; \frac{2V_f^2}{k+1}.$$

The left side of this desired inequality which we shall call e_k is built from the difference in the criterion values at $\hat{f}_k$ and an arbitrary f. It can be expressed as

$$e_k \;=\; \frac{1}{n}\sum_{i=1}^n[f(X_i)-\hat{f}_k(X_i)]+\log\int p_f(x)e^{\hat{f}_k(x)-f(x)} + \lambda[v_k-V_f],$$

where the integral arising from the ratio of the normalizers for $p_{\hat{f}_k}$ and p_f. Without loss of generality, making $\mathcal{H}$ closed under sign change, we restrict to positive β. This e_k is evaluated with $\hat{f}_k(x)=(1-\alpha)\hat{f}_{k-1}(x)+\beta h(x)$ and $v_k=(1-\alpha)v_{k-1}+\beta a_h$, at the optimized α,β and h, so we have that it is as least as good as at an arbitrary h with $\beta=\alpha v/a_h$ where $v=V_f$. Thus for any h we have that e_k is not more than

$$\frac{1}{n}\sum_{i=1}^n[f(X_i)-\bar{\alpha}\hat{f}_{k-1}(X_i)-\alpha h(X_i)/a_h]+$$

$$\log\int p_f(x)e^{[\bar{\alpha}\hat{f}_{k-1}(x)+\alpha v h(x)/a_h-f(x)]} \;+\; \bar{\alpha}\lambda[v_{k-1}-v],$$

where $\bar{\alpha}=(1-\alpha)$. Now reinterpret the integral using the expectation of $e^{\alpha[vh(x)/a_h-f(x)]}$ with respect to $p(x)=e^{\bar{\alpha}[f_{k-1}(x)-f(x)]}p_f(x)/c$, where c is its normalizing constant. Accordingly, we add and subtract $\log c=\log\int e^{\bar{\alpha}[f_{k-1}(x)-f(x)]}p_f(x)$ which, by Jensen's inequality using $\bar{\alpha}\le 1$, is not more than $\bar{\alpha}\log\int e^{[f_{k-1}(x)-f(x)]}p_f(x)$. Recognizing that this last integral is what arises in e_{k-1} and distributing f between the terms with coefficients $\bar{\alpha}$ and α, we obtain that e_k is not more than

$$\bar{\alpha}e_k+\alpha\frac{1}{n}\sum_{i=1}^n[f(X_i)-vh(X_i)/a_h]+\log\int e^{\alpha[vh(x)/a_h-f(x)]}p(x).$$

This inequality holds for all h so it holds in expectation with a random selection in which each h is drawn with probability $a_h|\theta_h|/v$ where the θ_h are the coefficients in the representation $f(x)=\sum_{h\in\mathcal{H}}\theta_h h(x)$ with $v=\sum_h|\theta_h|a_h=V_f$. We bring this expectation for random h inside the logarithm, and then inside the integral, obtaining an upper bound by Jensen's inequality. For each x and random h the quantities $[vh(x)/a_h-f(x)]$ have mean zero and have range of length not more than $2v$ since $a_h\ge\|h\|_\infty$. So by Hoeffding's moment generating function bound, the expectation for random h of $e^{\alpha[vh(x)/a_h-f(x)]}$ is not more than $e^{\alpha^2v^2/2}$. Thus

$$e_k \;\le\; (1-\alpha)e_{k-1}+\alpha^2V_f^2$$

for all $0\le\alpha\le 1$, and so in particular with $\alpha=2/(k+1)$. Also $e_0\le 2V_f^2$, so by induction

$$e_k\le\frac{2V_f^2}{k+1},$$

which is the desired result.

This computation bound and its regression counterpart in [13] is related to past relaxed greedy algorithm work (with $\lambda=0$ in [14], [2], [18], [9], [10], [17], [22], [3]. These previous results control the number of terms k rather than their ℓ_1 norm. The result stated here for ℓ_1 penalized log-likelihood and in [13] for regression, takes the matter a step further to show that with suitable positive λ the greedy pursuit algorithm solves the ℓ_1 penalized problem.

This computation analysis fits with our risk results. In the proof of Theorem 3.1, instead of the exact penalized likelihood estimator $\hat{f}$, substitute its k term greedy fit $\hat{f}_k$, The computation bound shows that this penalized likelihood ratio is not more than its corresponding value at any f, with addition of $2V_f^2/(k+1)$. Accordingly, its risk is not more than

$$Ed(f^*,\hat{f}_k)\;\le\;\min_{f\in\mathcal{F}}\left\{D(f^*,f)+\frac{\lambda_n V_f}{n}+\frac{2V_f^2}{k+1}\right\}+\frac{C}{n}.$$

The key step in our results is demonstration of approximation, computation, or covering properties, by showing that they hold on the average for certain distributions on the dictionary of possibilities.

REFERENCES

[1] A.R. Barron, "Complexity regularization with application to artificial neural networks," In G. Roussas (Ed.) *Nonparametric Functional Estimation and Related Topics*. pp.561–576. Dordrecht, the Netherlands, Kluwer Academic Publishers. 1990.

[2] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*. Vol. 39, pp.930–945. 1993.

[3] A.R. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.* Vol.36, pp.64–94. 2008.

[4] A.R. Barron and T.M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*. Vol.37, No.4, pp.1034–1054. 1991.

[5] A.R. Barron, C. Huang, J.Q. Li, X. Luo, "The MDL principle, penalized likelihoods, and statistical risk," In *Festschrift for Jorma Rissanen*. Tampere University Press, Tampere, Finland, 2008.

[6] A.R. Barron, C. Huang, J.Q. Li, X. Luo, "MDL, penalized likelihood and statistical risk," *IEEE Information Theory Workshop*. Porto Portugal, May 4-9, 2008.

[7] A.R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*. Vol.44, No.6, pp.2743–2760. 1998. Special Commemorative Issue: Information Theory: 1948-1998.

[8] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bull. Calcutta Math. Soc.* Vol.35, pp.99–109. 1943.

[9] G.H.L. Cheang *Neural Network Approximation and Estimation of Functions*. Ph.D. Thesis, Statistics Dept., Yale University. 1998.

[10] G.H.L. Cheang and A.R. Barron, "Penalized least squares, model selection, convex hull classes, and neural nets," In M. Verleysen (Ed.). *Proc. 9th ESANN*, pp.371–376. Brugge, Belgium, De-Facto Press. 2001.

[11] H. Cramér, *Mathematical Methods of Statistics*. Princeton Univ. Press. 1946.

[12] P. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, MIT Press. 2007.

[13] C. Huang, G.H.L. Cheang, and A.R. Barron. "Risk of penalized least squares, greedy selection and ℓ_1-penalization from flexible function libraries," Submitted to *Ann. Statist.* 2008.

[14] K.L. Jones, "A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.* Vol.20, pp.608–613. 1992.

[15] E.D. Kolaczyk and R.D. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Ann. Statist.* Vol.32, pp.500–527. 2004.

[16] J.Q. Li, *Estimation of Mixture Models*. Ph.D. Thesis, Statistics Dept., Yale University, 1999.

[17] J.Q. Li and A.R. Barron, "Mixture density estimation," In S. Solla, T. Leen, and K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems*, Vol.12, pp.279–285. 2000.

[18] W.S. Lee, P. Bartlett, and R.C. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in," *IEEE Trans. Inform. Theory*. Vol.42, pp.2118–2132. 1996.

[19] A. Rényi, "On measures of entropy and information," In *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol.1, pp.547–561. 1960.

[20] J. Rissanen, "Modeling by the shortest data description," *Automatica*. Vol.14, pp.465–471. 1978.

[21] J. Rissanen, "A universal prior on integers and estimation by minimum description length," *Ann. Statist.* Vol.11, pp.416–431. 1983.

[22] T. Zhang, "Sequential greedy approximation for certain convex optimization problems," *IEEE Trans. Inform. Theory*. Vol. 49, pp.682-691. 2003.