CHAPTER 13

Model Selection, Penalization and Oracle Inequalities

Our ultimate goal is to obtain sharper bounds on rates of convergence - in fact exactly optimal rates, rather than with spurious log terms.

However, this is a situation where the tools introduced are perhaps of independent interest. These include model selection via penalized least squares, where the penalty function is not ℓ_2 or even ℓ_1 but instead a function of the *number* of terms in the model. We will call such things *complexity penalties*.

Many of the arguments work for general (i.e. non-orthogonal) linear models. While we will not ultimately use this extra generality in this book, there are important applications (E.A.D.) and the model is of such importance that it seems reasonable to present part of the theory in this setting.

While it is natural to start with penalties proportional to the number of terms in the model, it will turn out that for our later results on exact rates, it will be necessary to consider a larger class of "$2k\log(p/k)$" penalties, in which, roughly speaking, the penalty to enter the k^{th} variable is a function that decreases with k approximately like $2\log(p/k)$.

We will be looking essentially at "all subsets" versions of the model selection problem. If there are p variables, then there are $\binom{p}{k}$ distinct submodels with k variables, and this grows very quickly with k. In order to control the resulting model explosion, good exponential probability inequalities for the tails of chi-square distributions are needed. We will derive these as a consequence of a powerful *concentration* inequality for Gaussian measures in $\mathbb{R}^n$. We give a separate exposition of this result, as it is finding increasing application in statistics.

13.1. A Gaussian concentration inequality

The Lipschitz norm of a function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\|f\|_{\mathrm{Lip}} = \sup\{\ |f(x) - f(y)|/\|x - y\|\ \}$$

where $\|x\|$ is the usual Euclidean norm on $\mathbb{R}^n$.

PROPOSITION 13.1. *If $Z \sim N_n(0, I)$, and $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz, then*

$$P\{f(Z) \geq Ef(Z) + t\} \leq e^{-t^2/(2\|f\|^2_{Lip})}. \tag{13.1}$$

Thus the tails of the distribution of a Lipschitz function of a Gaussian vector are sub Gaussian. Some statistically relevant examples of Lipschitz functions include

(i) order statistics. If $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(n)}$ denote the order statistics of a data vector z, then $f(z) = z_{(k)}$ has Lipschitz constant $\|f\|_{\mathrm{Lip}} = 1$.

(ii) ordered eigenvalues of symmetric matrices. Let A be an $n \times n$ symmetric matrix with eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$. If E is also symmetric, then

(e.g. (Golub & Van Loan 1996, p. 56 and 396))

$$|\lambda_k(A+E) - \lambda_k(A)| \le \|E\|_F,$$

where $\|E\|_F^2 = \sum_{i,j} e_{i,j}^2$ denotes the square of the *Frobenius* norm, which is the Euclidean norm on $n \times n$ matrices,

(iii) orthogonal projections. If S is a linear subspace of $\mathbb{R}^n$, then $f(z) = \|P_S z\|$ has Lipschitz constant 1. This is the example we use below. If $\dim S = k$, then $\|P_S z\|^2 \stackrel{\mathcal{D}}{=} \chi^2_{(k)}$ and so

$$E\|P_S z\| \le \{E\|P_S z\|^2\}^{1/2} = \sqrt{k}$$

and so the inequality implies

$$P\{\ \|P_S z\| \ge \sqrt{k} + t\ \} \le e^{-t^2/2}.$$

Note that the dimension n plays a very weak role in the inequality, which is sometimes said to be "infinite-dimensional". The phrase "concentration of measure" refers at least in part to the fact that the distribution of a Lipschitz(1) function of n variables is concentrated about its mean, in the sense that the tails are no heavier than those of a *univariate* standard Gaussian, regardless of the value of n! [1]

13.2. All subsets regression and complexity penalized least squares

We begin with the usual form of the general linear model with Gaussian errors:

$$y = X\beta + \epsilon z = \mu + \epsilon z, \qquad z \sim N_n(0, I). \tag{13.2}$$

There are n observations y and p unknown parameters β, connected by an $n \times p$ design matrix X with columns

$$X = [x_1, \cdots, x_p].$$

There is no restriction on p: indeed, we particularly wish to allow for situations in which $p \gg n$. We will assume that the noise level ϵ is known.

Example: Overcomplete dictionaries. Here is a brief indication of why one might wish to take $p \gg n$. Consider estimation of f in the continuous Gaussian white noise model (REF) $dY(t) = f(t)dt + \epsilon dW(t)$, and suppose that the observed data are inner products of Y with n orthonormal functions $\psi_1, \ldots, \psi_n$. Thus

$$y_i = \langle f, \psi_i \rangle + \epsilon z_i, \qquad i = 1, \ldots, n.$$

Now consider the possibility of approximating f by elements from a *dictionary* $\mathcal{D} = \{\phi_1, \phi_2, \ldots, \phi_p\}$. The hope is that by making $\mathcal{D}$ sufficiently rich, one might be able to represent f well by a linear combination of a very few elements of $\mathcal{D}$. This idea has been advanced by a number of authors `ADD REFERENCES.` As a simple illustration, the ψ_i might be sinusoids at the first n frequencies, while the dictionary elements might allow a much finer sampling of frequencies

$$\phi_k(t) = \sin(2\pi k t/p), \qquad k = 1, \ldots, p = n^\beta \gg n.$$

with $p = n^\beta$ for some $\beta > 1$. If there is a single dominant frequency in the data, it is possible that it will be essentially captured by an element of the dictionary even if it does not complete an integer number of cycles in the sampling interval.

[1] In fact, sharper bounds for the tail of χ^2 random variables are available (Laurent & Massart (1998), Johnstone (2001), Birgé & Massart (2001), `[CHECK!]`), but this bound will suffice for our purposes and serves as an illustration of the power of the general inequality (13.1).

If we suppose that f has the form $f = \sum_{j=1}^{p} \beta_j \phi_j$, then these observation equation become an instance of the general linear model (13.2) with

$$X_{ij} = \langle \psi_i, \phi_j \rangle.$$

Again, the hope is that one can find an estimate $\hat{\beta}$ for which only a small number of components $\hat{\beta}_j \neq 0$.

All subsets regression. To each subset $J \subset \{1, \ldots, p\}$ of cardinality $n_J = |J|$ corresponds a regression model which fits only the variables x_j for $j \in J$. The possible fitted vectors μ that could arise from these variables lie in the model space

$$S_J = \text{span}\{x_j : j \in J\}.$$

The dimension of S_J is at most n_J, and could be less in the case of collinearity.

Let P_J denote orthogonal projection onto S_J: the least squares estimator $\hat{\mu}_J$ of μ is given by $\hat{\mu}_J = P_J y$. The issue in all subsets regression consists in deciding how to select a subset $\hat{J}$ on the basis of data y: the resulting estimate of μ is then $\hat{\mu} = P_{\hat{J}} y$.

Mean squared error properties can be used to motivate all subsets regression. We will use a predictive risk[2] criterion to judge an estimator $\hat{\beta}$ through the fit $\hat{\mu} = X\hat{\beta}$ that it generates:

$$E\|X\hat{\beta} - X\beta\|^2 = E\|\hat{\mu} - \mu\|^2.$$

The mean of a projection estimator $\hat{\mu}_J$ is just the projection of μ, namely $E\hat{\mu}_J = P_J\mu$, while its variance is $\epsilon^2 \text{tr} P_J = \epsilon^2 \dim S_J$. From the variance-bias decomposition of MSE,

$$E\|\hat{\mu}_J - \mu\|^2 = \|P_J\mu - \mu\|^2 + \epsilon^2 \dim S_J.$$

A *saturated* model arises from any subset with $\dim S_J = n$, so that $\hat{\mu}_J = y$ "interpolates the data". In this case the MSE is just the unrestricted minimax risk for $\mathbb{R}^n$:

$$E\|\hat{\mu} - \mu\|^2 = n\epsilon^2.$$

Comparing the last two displays, we see that if μ lies close to a low rank subspace — $\mu \doteq \sum_{j \in J} \beta_j x_j$ for $|J|$ small—then $\hat{\mu}_J$ offers substantial risk savings over a saturated model. Thus, it seems that one would wish to expand the dictionary $\mathcal{D}$ as much as possible to increase the possibilities for sparse representation. Against this must be set the dangers inherent in fitting over-parametrized models – principally overfitting of the data. Penalized least squares estimators are designed specifically to address this tradeoff.

This discussion leads to a natural generalization of the notion of ideal risk introduced in Chapter ???. For each mean vector μ, there will be an optimal model subset $J = J(\mu)$ which attains the ideal risk

$$\mathcal{R}(\mu, \epsilon) = \min_J \ \|\mu - P_J\mu\|^2 + \epsilon^2 \dim S_J.$$

[2]Why the name "predictive risk"? Imagine that new data will be taken from the same design as used to generate the original observations y and estimator $\hat{\beta}$: $y^* = X\beta + \epsilon z^*$. A natural prediction of y^* is $X\hat{\beta}$, and its mean squared error, averaging over the distributions of both z and z^*, is

$$E\|y^* - X\hat{\beta}\|^2 = E\|X\beta - X\hat{\beta}\|^2 + n\epsilon^2,$$

so that the mean squared error of prediction equals $E\|\hat{\mu} - \mu\|^2$, up to an additive factor that doesn't depend on the model chosen.]

Of course, this choice $J(\mu)$ is not available to the statistician, since μ is unknown. The challenge, taken up below, is to see to what extent penalized least squares estimators can "mimick" ideal risk, in a fashion analagous to the mimicking achieved by threshold estimators in the orthogonal setting.

Complexity penalized least squares The residual sum of squares (RSS) after fitting model J is

$$\|y - \hat{\mu}_J\|^2 = \|y - P_J y\|^2,$$

and clearly decreases as the model J increases. To discourage simply using a saturated model, we introduce a penalty on the size of the model, $\mathrm{pen}(n_J)$, and define a complexity criterion

$$C(J, y) = \|y - \hat{\mu}_J\|^2 + \epsilon^2 \mathrm{pen}(n_J). \tag{13.3}$$

The complexity penalized RSS estimate $\hat{\mu}_{\mathrm{pen}}$ is then given by orthogonal projection onto the subset that minimizes the penalized criterion:

$$\begin{aligned} \hat{J}_{\mathrm{pen}} &= \mathrm{argmin}_J \; C(J, y) \\ \hat{\mu}_{\mathrm{pen}} &= P_{\hat{J}_{\mathrm{pen}}} y. \end{aligned} \tag{13.4}$$

The simplest penalty function is simply proportional to the number of variables in the model:

$$\mathrm{pen}_0(k) = \lambda_p^2 \, k, \tag{13.5}$$

where we will take λ_p^2 to be roughly of order $2 \log p$. [The well known AIC criterion would set $\lambda_p^2 = 2$: this is effective for selection among a nested sequence of models, but is known to overfit in all-subsets settings. `[mention BIC` $\lambda_p^2 = \log p$`?]`

For this particular case, we describe the kind of oracle inequality to be proved in this chapter. First, note that for $\mathrm{pen}_0(k)$, minimal complexity and ideal risk are related:

$$\begin{aligned} \min_J C(J, \mu) &= \min_J \; [\, \|\mu - P_J \mu\|^2 + \epsilon^2 \mathrm{pen}_0(n_J)] \\ &\le \lambda_p^2 \min \, [\, \|\mu - P_J \mu\|^2 + \epsilon^2 n_J] = \lambda_p^2 \; \mathcal{R}(\mu, \epsilon). \end{aligned}$$

Let $\lambda_p = \zeta(1 + \sqrt{2 \log p})$ for $\zeta > 1$ and $A(\zeta) = (1 - \zeta^{-1})^{-1}$. Then for penalty function (13.5) and arbitary μ,

$$E\|\hat{\mu}_{\mathrm{pen}} - \mu\|^2 \le A(\zeta) \lambda_p^2 [C\epsilon^2 + \mathcal{R}(\mu, \epsilon)].$$

Thus, the complexity penalized RSS estimator, for non-orthogonal and possibly over-complete dictionaries, comes within a factor of order $2 \log p$ of the ideal risk.

Remark. Another possibility is to use penalty functions monotone in the rank of the model, $\mathrm{pen}(\dim J)$. However, when $k \to \mathrm{pen}(k)$ is strictly monotone, this will yield the same models as minimizing (13.3), since a collinear model will always be rejected in favor of a sub-model with the same span.

13.3. Oracle inequalities for $2k\log(p/k)$ penalties

Class $\mathcal{P}$. Assume that $k \to \mathrm{pen}(k)$ is strictly increasing, and has the specific form

$$\mathrm{pen}(k) = \zeta^2 k(1 + \sqrt{2L_{p,k}})^2 \qquad (\zeta > 1)$$

with $k \to L_{p,k}$ non-increasing and

$$L_{p,k} \geq \log(p/k) + \gamma_k$$

for some sequence γ_k not depending on p and satisfying $\gamma_k \geq \gamma > 1$ for all large k.

The class $\mathcal{P}$ is broad enough to include penalties proportional to model size, c.f. (13.5), with $L_{p,k} \equiv \lambda_p \geq \log p$. However, the possibility that $k \to L_{p,k}$ be strictly decreasing is critical for the later application of the oracle inequalities to derive exact rates of convergence. Since the dominant terms in the two preceding displays are $2k\log(p/k)$, we will loosely refer to $(\mathcal{P})$ as a class of "$2k\log(p/k)$ penalties".

THEOREM 13.2. *Let $\hat{\mu}$ be a penalized least squares estimate of the form (13.3)-(13.4) for a penalty of class $\mathcal{P}$ defined above. Then, for all μ,*

$$E\|\hat{\mu} - \mu\|^2 \leq (1 - \zeta^{-1})^{-1}\big[\zeta C_\gamma L_{p,1}\epsilon^2 + \min_J C(J,\mu)\big].$$

We will see that $\min_J C(J,\mu)$ is an extension of the notion of ideal risk, and describes the "intrinsic accuracy" of approximation of μ by models $\{S_J : J \subset \{1,\ldots,p\}\}$.

Our key examples will have the form $L_{p,k} \approx \log(p/k) + \gamma$, for $\gamma > 1$. For example, in the orthogonal case, the False Discovery Rate thresholds satisfy

$$t_{n,k} \sim \sqrt{2\log(2n/kq)}$$

corresponding roughly to $L_{p,k} = \log(n/k) + \gamma$ with $\gamma = \log(2/q) > 1$ for $q < 2/e$.

"Threshold" interpretation. In the orthogonal setting ($n = p$), suppose that

$$\mathrm{pen}(k) = \sum_{j=1}^{k} t_{n,j}^2,$$

so that $\hat{\mu}_{\mathrm{pen}}$ becomes thresholding:

$$\hat{\mu}_{\mathrm{pen},j} = \begin{cases} y_j & |y_j| \geq \epsilon t_{n,\hat{k}}, \\ 0 & \text{otherwise} \end{cases}$$

where

$$\hat{k} = \operatorname*{argmin}_{0\leq k\leq n} \sum_{k+1}^{n} y_{(j)}^2 + \epsilon^2 \sum_{1}^{k} t_{p,j}^2.$$

Typically $\sum_1^k t_{n,j}^2 \sim k t_{n,k}^2$, at least for $k = o(n)$, and so the thresholds covered by the Theorem have the form

$$t_{n,k} \approx \zeta(1 + \sqrt{2L_{p,k}}) \geq \zeta(1 + \sqrt{2\log(p/k) + 2\gamma_n}).$$

The fact that $\zeta > 1$ (rather than $= 1$) and the extra "1" are needed for the proof, but do make the thresholds somewhat larger than is desirable in practice.

The idea to use penalties of the general form $2\epsilon^2 k \log(n/k)$ arose among several authors more or less simultaneously:

- Foster & Stine (1997) $\text{pen}(k) = \epsilon^2 \sum_1^k 2\log(n/j)$ via information theory.
- George & Foster (2000) Empirical Bayes approach. $[\mu_i \overset{i.i.d.}{\sim} (1-w)\delta_0 + wN(0,C)$ followed by estimation of $(w,C)]$. They argue that this approach penalizes the k^{th} variable by about $2\epsilon^2 \log(((n+1)/k) - 1)$.
- The covariance inflation criterion of Tibshirani & Knight (1999) in the orthogonal case leads to $\text{pen}(k) = 2\epsilon^2 \sum_1^k 2\log(n/j)$.
- FDR - discussed above (?).
- Birgé & Massart (2001) contains a systematic study of complexity penalized model selection from the specific viewpoint of obtaining non-asymptotic bounds, using a penalty class similar to, but more general than that used here.

13.4. Proof of Oracle inequality

Penalized complexity and complexity functionals The definition of the complexity penalized RSS estimator uses the minimum over submodels J of

$$C(J,y) = \|y - P_J y\|^2 + \text{pen}(n_J).$$

It will be helpful to introduce a related *complexity* functional $K(\mu, y)$, defined for all $\mu \in \mathbb{R}^n$. First, a definition. Given μ, the minimal dimension of a model containing μ is

$$N(\mu) = \inf\{n_J : \mu \in S_J\}.$$

Now define

$$K(\mu,y) = \|y-\mu\|^2 + \text{pen}(N(\mu)).$$

The criteria C and K yield the same estimators:

LEMMA 13.3. *Suppose that pen(k) is strictly increasing in k. Given data y, let $\hat{J}$ be a minimizer of $C(J,y)$ and set $\hat{\mu} = P_{\hat{J}} y$. Then $\hat{\mu}$ minimizes $K(\mu,y)$ and*

$$\inf_\mu K(\mu,y) = \min_J C(J,y).$$

PROOF. We chase definitions resolutely. First, fix μ and let $J(\mu) = \text{argmin}\{n_J : \mu \in S_J\}$. Consequently $\mu \in S_{J(\mu)}$ and hence both

$$\|y-\mu\|^2 \geq \|y - P_{J(\mu)} y\|^2 \qquad \text{and} \qquad N(\mu) = n_{J(\mu)},$$

so that for every μ,

$$K(\mu,y) \geq C(J(\mu),y) \geq \min_J C(J,y).$$

Now turn to $\hat{J} = \text{argmin}\, C(J,y)$ and $\hat{\mu} = P_{\hat{J}} y$. Let us show that $N(\hat{\mu}) = n_{\hat{J}}$. Indeed, suppose to the contrary that there were a subset J with $n_J < n_{\hat{J}}$ and $\hat{\mu} \in S_J$. Then we would have both $\text{pen}(n_J) < \text{pen}(n_{\hat{J}})$ and

$$\|y - P_J y\|^2 \leq \|y - \hat{\mu}\|^2 = \|y - P_{\hat{J}} y\|^2.$$

But this means that $C(J,y) < C(\hat{J},y)$, in contradiction to the definition of $\hat{J}$. Hence $\text{pen}(N(\hat{\mu})) = \text{pen}(n_{\hat{J}})$ and so

$$K(\hat{\mu},y) = \|y - P_{\hat{J}} y\|^2 + \text{pen}(n_{\hat{J}}) = C(\hat{J},y) = \min_J C(J,y).$$

Combining the second and fourth displays, we obtain the result. □

It is now clear that Theorem 13.2 is a consequence of

THEOREM 13.4. *Let $K(\mu, y) = \|y - \mu\|^2 + \epsilon^2 pen(N(\mu))$ and assume that pen(k) satsifies assumptions $\mathcal{P}$. If $\hat{\mu}_P = \operatorname{argmin}_{\tilde{\mu}} K(\tilde{\mu}, y)$ and $K_0(\mu) = \inf_{\tilde{\mu}} K(\tilde{\mu}, \mu)$, then*

$$EK(\hat{\mu}_P, \mu) \le (1 - \zeta^{-1})^{-1}\big[\zeta C_\gamma L_{p,1}\epsilon^2 + K_0(\mu)\big].$$

PROOF. 1°. Reduction to $\epsilon = 1$. Showing the $\epsilon-$dependence explicitly for now in $K_\epsilon(\mu, y)$, we see immediately that

$$K_\epsilon(\mu, y) = \epsilon^2 K_1(\mu/\epsilon, y/\epsilon),$$

and so it suffices to establish the inequality when $\epsilon = 1$.

2°. *A basic inequality* One reason to introduce the functional $K(\mu, y)$ is a useful inequality which we now derive. Indeed, use of $y = \mu + \epsilon z$ and the expansion $\|y - \tilde{\mu}\|^2 = \|\epsilon z\|^2 + 2\langle \mu - \tilde{\mu}, \epsilon z\rangle + \|\mu - \tilde{\mu}\|^2$ lead to the identity

$$K(\tilde{\mu}, y) = \|\epsilon z\|^2 + 2\langle \mu - \tilde{\mu}, \epsilon z\rangle + K(\tilde{\mu}, \mu).$$

Use this identity for both $K(\hat{\mu}, y)$ and $K(\tilde{\mu}, y)$. Since $K(\hat{\mu}, y) \le K(\tilde{\mu}, y)$ for all $\tilde{\mu}$ by definition, we obtain by subtraction the basic inequality

$$K(\hat{\mu}, \mu) \le K(\tilde{\mu}, \mu) + 2\langle \hat{\mu} - \tilde{\mu}, \epsilon z\rangle. \tag{13.6}$$

The left side exceeds the quadratic loss $\|\hat{\mu} - \mu\|^2$, while on the right side, $\tilde{\mu}$ can be chosen to minimize the *theoretical complexity* $K_0(\mu) = \inf_{\tilde{\mu}} K(\tilde{\mu}, \mu)$. If a minimizing value is μ_0, the chief task in obtaining an upper bound for $E\|\hat{\mu} - \mu\|^2$ in terms of $K_0(\mu)$ becomes that of bounding the error term $\langle \hat{\mu} - \mu_0, \epsilon z\rangle$.

We return to the basic inequality (13.6). Inserting μ_0 for $\tilde{\mu}$ and noting that $E\langle \mu - \mu_0, z\rangle = 0$, we have

$$EK(\hat{\mu}, \mu) \le K(\mu_0, \mu) + 2E\langle \hat{\mu} - \mu, z\rangle. \tag{13.7}$$

Our goal will be to derive an "a priori" bound for $E\langle \hat{\mu} - \mu, z\rangle$ in terms of $\zeta^{-1}EK(\hat{\mu}, \mu)$ and an error term. This will require control on deviations of $\|P_S z\|^2$ as S ranges over the possible subset models.

3°. *Rejection of individual models.* Given a model S, we will say that it is rejected if

$$\|P_S z\|^2 \ge \zeta^{-2}\mathrm{pen}(\dim S). \tag{13.8}$$

In a 'null hypothesis' situation, this will constitute a type I error, and so the penalty should be large enough so as to make this type I error probability small, simultaneously over all sufficiently large models S.

With the choice $\mathrm{pen}(k) = \zeta^2 k(1 + \sqrt{2L_{p,k}})^2$, we find that when $\dim S = k$, (13.8) is equivalent to

$$\|P_S z\| \ge \sqrt{k} + \sqrt{2kL_{p,k}}.$$

Now $E\|P_S z\|^2 = E\chi^2_{(k)} = k$, and so $E\|P_S z\| \le \sqrt{k}$. Consequently, from the concentration inequality (13.1),

$$P\{\text{reject model } S\} \le e^{-kL_{p,k}}. \tag{13.9}$$

4°. *A Priori Bound.* Fix μ, and for each model J, let $S_{J,\mu} = \mathrm{span}\{\mu, S_J\}$. Denote the full augmented collection of models by

$$\mathcal{M}_\mu = \{S_J, S_{J,\mu} : J \subset 1 : p\}.$$

We need to track the size the of the largest rejected model. So, let

$$\hat{N} = \begin{cases} \min\{N : & \text{all models } S \in \mathcal{M}_\mu \text{ with } \dim S \geq n \text{ are "accepted"}\} \\ \infty & \text{if } \|z\|^2 \geq \zeta^{-2}\text{pen}(n), \end{cases}$$

since in the latter case, even the saturated model is not accepted. Correspondingly, set

$$\hat{P}(z,\mu) = \begin{cases} \text{pen}(\hat{N}) & \text{if } \hat{N} \leq n \\ \|\zeta z\|^2 & \text{if } \hat{N} = \infty. \end{cases}$$

We will establish that

$$2\zeta\langle z, \mu' - \mu\rangle \leq \hat{P}(z,\mu) + K(\mu',\mu). \tag{13.10}$$

Assuming for now, both (13.10) and that $E\hat{P} < \infty$, we have, on substitution into (13.7),

$$EK(\hat{\mu},\mu) \leq K_0(\mu) + \zeta^{-1}E_\mu\hat{P} + \zeta^{-1}EK(\hat{\mu},\mu).$$

Moving the unknown to the left side and then ignoring the penalty term in $K(\hat{\mu},\mu) = \|\hat{\mu} - \mu\|^2 + \text{pen}(\hat{\mu})$, we get

$$E\|\hat{\mu} - \mu\|^2 \leq (1 - \zeta^{-1})^{-1}\big[\zeta^{-1}E\hat{P} + K_0(\mu)\big].$$

5°. *Proof of (13.10).* As before, let $J(\mu') = \text{argmin}\{n_J : \mu' \in S_J\}$ be the minimal model containing μ', and set

$$S' = S_{J(\mu'),\mu} = \text{span}\{\mu, S_{J(\mu')}\}.$$

Clearly $\mu' - \mu \in S'$ and so

$$\begin{aligned} 2\zeta\langle z, \mu' - \mu\rangle &\leq 2\zeta\|P_{S'}z\|\,\|\mu' - \mu\| \\ &\leq \|\zeta P_{S'}z\|^2 + \|\mu' - \mu\|^2. \end{aligned}$$

If $\hat{N} = \infty$, the right side is bounded by $\|\zeta z\|^2 + K(\mu',\mu)$, so the Claim is straightforward. So we now suppose that $\hat{N} \leq n$, and propose to show that

$$\|\zeta P_{S'}z\|^2 \leq \text{pen}(\hat{N}) + \text{pen}(N(\mu')),$$

which suffices for (13.10).

Two cases arise. If $\dim S' \leq \hat{N}$, choose $S'' \supset S'$ with $\dim S'' = \hat{N}$. By definition of $\hat{N}$

$$\|\zeta P_{S'}z\|^2 \leq \|\zeta P_{S''}z\|^2 \leq \text{pen}(\hat{N}).$$

In the second case, $\dim S' > \hat{N}$. Since $N(\mu') \leq \dim S' \leq N(\mu') + 1$, we know also that $N(\mu') \geq \hat{N}$. We remark in addition that for any r, $\text{pen}(r+1) \leq \text{pen}(r) + \text{pen}(r)/r$ since $\text{pen}(r)/r$ is decreasing. Since model S' is accepted, we therefore find

$$\begin{aligned} \|\zeta P_{S'}z\|^2 &\leq \text{pen}(\dim S') \\ &\leq \text{pen}(N(\mu')) + \text{pen}(N(\mu'))/N(\mu') \\ &\leq \text{pen}(N(\mu')) + \text{pen}(\hat{N})/\hat{N} \\ &\leq \text{pen}(N(\mu')) + \text{pen}(\hat{N}). \end{aligned}$$

6°. *Bounding* $E\hat{P}$. By definition, we have

$$E\hat{P}(z,\mu) = E\{\text{pen}(\hat{N}), \hat{N} \le n\} + E\{\|\zeta z\|^2, \|\zeta z\|^2 \ge \text{pen}(n)\}, \tag{13.11}$$

and, using monotonicity of $k \to L_{p,k}$ and setting $\lambda_{p,1} = \zeta(1 + \sqrt{2L_{p,1}})$,

$$E\{\text{pen}(\hat{N}), \hat{N} \le n\} \le \lambda_{p,1}^2 \{P(\hat{N} = 1) + \sum_{k=1}^{n-1} (k+1) P(\hat{N} = k+1)\}.$$

If $\hat{N} = k+1$, there is some $k-$ dimensional model in $\mathcal{M}_\mu$ that is rejected. For each individual model, the rejection probability is bounded by (13.9). We turn to bounding the number of such models.

Let $\mathcal{S}_k$ be the set of models S_J with dimension k—clearly $|(\mathcal{S}_k)| \le \binom{p}{k}$. A model in $\mathcal{M}_\mu$ of dimension k is either from $\mathcal{S}_k$ or is of the form $S_J \oplus \mu$, for some $S_J \in \mathcal{S}_{k-1}$ which does not already contain μ. Hence

$$\begin{aligned} |\{S \in \mathcal{M}_\mu : \dim S = k\}| &\le |\mathcal{S}_k| + |\mathcal{S}_{k-1}| \\ &\le \binom{p}{k} + \binom{p}{k-1} \le 2\frac{p^k}{k!} \end{aligned}$$

since $k < n \le p$. Putting this together with (13.9) gives

$$\begin{aligned} P\{\hat{N} = k+1\} &\le 2\frac{p^k}{k!} \exp\{-kL_{p,k}\} \\ &\le \frac{2}{\sqrt{2\pi k}} \exp\{-k[L_{p,k} - \log(p/k) - 1]\}, \end{aligned}$$

where we have used Stirling's inequality: $k! \ge \sqrt{2\pi} e^{-k} k^{k+1/2}$.

Now use the assumption $L_{p,k} \ge \log(p/k) + \gamma$ for some $\gamma > 1$. This yields

$$E[\hat{N}, \hat{N} \le n] \le 1 + \sum_{k=1}^{n-1} \frac{2(k+1)}{\sqrt{2\pi k}} e^{-k(\gamma-1)} \le c(\gamma).$$

We must finally deal with the second term in (13.11). The Cauchy-Schwarz inequality shows it to be bounded by

$$\{E\|\zeta z\|^4\}^{1/2} \; P\{\|z\| > \sqrt{n} + \sqrt{2nL_{p,n}}\}^{1/2}.$$

Since $\|z\|^2 \sim \chi_n^2$, we have $E\|z\|^4 = 2n + n^2 \le (n+1)^2$. Applying (13.1) to the second term, we obtain as upper bound

$$\zeta^2(n+1) e^{-\frac{1}{2}nL_{p,n}} \le \zeta^2(n+1) e^{-n/2},$$

using again the assumption on $L_{p,k}$.

Finally assembling the bounds, we have, so long as $L_{p,1} \ge 1$,

$$\begin{aligned} EP(z,\mu) &\le \zeta^2(1 + \sqrt{2L_{p,1}})^2 c(\gamma) + \zeta^2(n+1)e^{-n/2} \\ &\le \zeta^2(1 + \sqrt{2L_{p,1}})^2 c'(\gamma) \le \zeta^2 c''(\gamma) L_{p,1}. \end{aligned}$$

□

13.5. Aside: Stepwise methods vs. complexity penalization.

Stepwise model selection methods have long been used as heuristic tools for model selection. In this aside, we explain a connection between such methods and a class of penalties for penalized least squares.

The basic idea with stepwise methods is to use a test statistic—in application, often an F-test—and a threshold to decide whether to add or delete a variable from the current fitted model. Let $\hat{J}_k$ denote the best submodel of size k:

$$\hat{J}_k = \text{argmax}_k\{\|P_J y\|^2 : n_J = k\},$$

and denote the resulting best k-variable estimator by $Q_k y = P_{\hat{J}_k} y$. The mapping $y \to Q_k(y)$ is non-linear since the optimal set $\hat{J}_k(y)$ will in general vary with y.

In the *forward stepwise* approach, the model size is progressively increased until a threshold criterion suggests that no further benefit will accrue by continuing. Thus, define

$$\hat{k}_G = \text{first } k \text{ s.t. } \|Q_{k+1}y\|^2 - \|Q_k y\|^2 \le \epsilon^2 t^2_{p,k+1}. \tag{13.12}$$

Note that we allow the threshold to depend on k: in practice it is often constant, but we wish to allow $k \to t^2_{p,k}$ to be decreasing.

In contrast, the *backward stepwise* approach starts with a saturated model and gradually decreases model size until there appears to be no further advantage in going on. So, define

$$\hat{k}_F = \text{last } k \text{ s.t. } \|Q_k y\|^2 - \|Q_{k-1}y\|^2 \ge \epsilon^2 t^2_{p,k}. \tag{13.13}$$

Remarks. 1. In the orthogonal case, $y_i = \mu_i + \epsilon z_i, i = 1, \dots, n$ with order statistics $|y|_{(1)} \ge |y|_{(2)} \ge \dots \ge |y|_{(n)}$, we find that

$$\|Q_k y\|^2 = \sum_{j=1}^{k} |y|^2_{(j)},$$

so that

$$\hat{k}_F = \max\{k : |y|_{(k)} \ge \epsilon t_{p,k}\}, \tag{13.14}$$

and that $\hat{k}_F$ agrees with the FDR definition with $t_{p,k} = z(qk/2n)$. `[Fuller reference??]` In this case, it is critical to the method that the thresholds $k \to t_{p,k}$ be (slowly) decreasing.

2. In practice, for reasons of computational simplicity, the forward and backward stepwise algorithms are often "greedy", i.e., they look for the best variable to add (or delete) without optimizing over all sets of size k.

The stepwise schemes are related to a penalized least squares estimator. Let

$$\begin{aligned} S(k) &= \|y - Q_k y\|^2 + \epsilon^2 \sum_{j=1}^{k} t^2_{p,j}, \\ \hat{k}_2 &= \text{argmin}_{0 \le k \le n}\ S(k). \end{aligned} \tag{13.15}$$

Thus the associated penalty function is $\text{pen}(k) = \sum_1^k t^2_{p,j}$ and the corresponding estimator is given by (13.3) and (13.4). `[Remark that` $pen(k)$ `satisfies`

assumptions (P) if $j \to t_{p,j}$ is decreasing with $\zeta^{-1} t_{p,k} \geq \log(p/k) + \gamma_k$ etc. ??]

The optimal model size for pen(k) is bracketed between the stepwise quantities.

PROPOSITION 13.5. *Let $\hat{k}_G, \hat{k}_F$ be the forward and backward stepwise variable numbers defined at (13.12) and (13.13) respectively, and let $\hat{k}_2$ be the global optimum model size for pen(k) defined at (13.15). Then*

$$\hat{k}_G \leq \hat{k}_2 \leq \hat{k}_F.$$

PROOF. Since $\|y - Q_k y\|^2 = \|y\|^2 - \|Q_k y\|^2$,

$$S(k+1) - S(k) = \|Q_k y\|^2 - \|Q_{k+1} y\|^2 + \epsilon^2 t_{p,k+1}^2.$$

Thus

$$S(k+1) \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} S(k) \quad \text{according as} \quad \|Q_{k+1} y\|^2 - \|Q_k y\|^2 \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} \epsilon^2 t_{p,k+1}^2.$$

Thus, if it were the case that $\hat{k}_2 > \hat{k}_F$, then necessarily $S(\hat{k}_2) > S(\hat{k}_2 - 1)$, which would contradict the definition of $\hat{k}_2$ as a global minimum of $S(k)$. Likewise, $\hat{k}_2 < \hat{k}_G$ is not possible, since it would imply that $S(\hat{k}_2 + 1) < S(\hat{k}_2)$. □

[Include diagram, notes 11/26/02 p.8 showing local and global minima.]

[Remark?: Experience with FDR in orthogonal case shows that often $\hat{k}_G = \hat{k}_2 = \hat{k}_F$, and that in sparse cases can bound $\hat{k}_F - \hat{k}_G$.]

13.6. A version for orthogonal regression

We now specialize to the $n-$ dimensional white Gaussian sequence model: [3]

$$y_i = \mu_i + \epsilon z_i, \qquad i = 1, \ldots, n, \qquad z_i \overset{i.i.d.}{\sim} N(0,1). \tag{13.16}$$

The columns of the design matrix implicit in (13.16) are the unit co-ordinate vectors e_i, consisting of zeros except for a 1 in the i^{th} position. The least squares estimator corresponding to a subset $J \subset \{1, \ldots, n\}$ is simply given by co-ordinate projection:

$$\hat{\mu}_{J,k}(y) = \begin{cases} y_k & k \in J \\ 0 & k \notin J. \end{cases}$$

In the orthogonal setting

$$N(\mu) = \#\{i \ : \ \mu_i \neq 0\}$$

simply counts the number of non-zero co-ordinates. To minimize the empirical complexity

$$K(\mu, y) = \|y - \mu\|^2 + \epsilon^2 \text{pen}(N(\mu)),$$

[3] Of course, this is the canonical form of the more general orthogonal regression setting

$$Y = X\beta + \epsilon Z,$$

with N dimensional response and n dimensional parameter vector β linked by an orthogonal design matrix X satisfying $X^T X = I_n$, and with the noise $Z \sim N_N(0, I)$. This reduces to (13.16) after premultiplying by X^T and setting $y = X^T Y$, $\mu = \beta$ and $z = X^T Z$.

we observe that the best fitting model of dimension k just picks off the largest k observations. Thus, if $y_{(1)}^2 \geq y_{(2)}^2 \geq \ldots \geq y_{(n)}^2$ denote the ordered squared observations,

$$\begin{aligned}\inf_{\mu} K(\mu, y) &= \min_k \inf_{N(\mu)=k} \quad \|y-\mu\|^2 + \epsilon^2 \mathrm{pen}(k) \\ &= \min_k \quad \sum_{j>k} y_{(j)}^2 + \epsilon^2 \mathrm{pen}(k).\end{aligned}$$

If $\hat{k}$ denotes the (data dependent!) minimizing value of k, then the complexity penalized least squares estimate $\hat{\mu}_P$ is given by hard thresholding at threshold $\hat{t} = |y|_{(\hat{k})}$:

$$\hat{\mu}_{P,j}(y) = \begin{cases} y_j & \text{if } |y_j| \geq \hat{t} \\ 0 & \text{otherwise.} \end{cases}$$

To state the oracle inequality established in the last section (?), recall the definition of theoretical complexity:

$$K_\epsilon(\mu) = \inf_{\tilde{\mu}} \|\mu - \tilde{\mu}\| + \epsilon^2 \mathrm{pen}(N(\tilde{\mu})),$$

and the assumptions (P) on the penalty function, namely that

$$\mathrm{pen}(k) = \zeta^2 k(1 + \sqrt{2L_{n,k}})^2, \qquad \zeta > 1,$$

and (i) $k \to \mathrm{pen}(k)$ is strictly increasing, (ii) $k \to L_{n,k}$ is nonincreasing, and (iii) $L_{n,k} \geq \log(n/k) + \gamma_k$, with $\gamma_k \geq \gamma > 1$ for $k \geq k_0$.

Then the oracle inequality becomes

$$E\|\hat{\mu}_P - \mu\|^2 \leq c_1 L_{n,1} \epsilon^2 + c_2 K_\epsilon(\mu), \tag{13.17}$$

with $c_1 = c_1(\zeta, \gamma)$ and $c_2 = c_2(\zeta)$.

A simple bound for theoretical complexity

LEMMA 13.6. *If pen$(k) = k\lambda_k^2$ with $k \to \lambda_k$ non-increasing, then*

$$K_\epsilon(\mu) \leq \sum_{k=1}^n \mu_{(k)}^2 \wedge \lambda_k^2 \epsilon^2. \tag{13.18}$$

This and the following lemma will be proved below. For now, look at the quantity

$$\mathcal{R}_\lambda(\mu, \epsilon) = \sum_{k=1}^n \mu_{(k)}^2 \wedge \lambda_k^2 \epsilon^2.$$

If $\lambda_k \equiv \lambda$, this reduces to the ideal risk $\mathcal{R}(\mu, \lambda\epsilon)$ studied at length earlier – and in fact, equality holds in (13.18) in this case. However, we are now more interested in cases in which λ_k is strictly decreasing, for example like $k \to \zeta(1 + \sqrt{2\log(n\beta/k)})$. Although the rate of decrease is slow, we will see that it suffices to remove spurious logarithmic terms from rates of convergence.

We remark also that in typical cases, the inequality in (13.18) is sharp at the level of rates:

LEMMA 13.7. *Suppose that $\lambda_k = \ell(k/n)$ for a function $\ell(x)$, positive and decreasing in $x \in [0,1]$, that satisfies*

$$\lim_{x\to 0} x\ell(x) = 0, \qquad \sup_{0\le x\le 1} x\ell'(x) \le c_1.$$

Then

$$\sum_{k=1}^{n} \mu_{(k)}^2 \wedge \lambda_k^2\epsilon^2 \le c_2 K_\epsilon(\mu), \qquad c_2 \le 1 + c_1/\ell(1).$$

This leads immediately to an important bound.

COROLLARY 13.8. *If $\hat\mu_P = argmin_\mu \|y-\mu\|^2 + \epsilon^2 pen(N(\mu))$, for $pen(k) = k\lambda_k^2$ and $\lambda_k = \zeta(1+\sqrt{2L_{n,k}})$ satisfying assumptions (P), then*

$$E\|\hat\mu_P - \mu\|^2 \le c_1 L_{n,1}\epsilon^2 + c_2 \sum_k \mu_{(k)}^2 \wedge \lambda_k^2\epsilon^2.$$

`[Include proofs, pp 5 & 6.]`

LEMMA 13.9. *(a) Suppose that $\{s_k\}_1^n$ and $\{\gamma_k\}_1^n$ are positive, non-decreasing sequences. Then*

$$\min_{0\le k\le n} \; [ks_k + \sum_{k+1}^{n} \gamma_n] \;\le\; \sum_{k=1}^{n} s_k \wedge \gamma_k.$$

(b) Conversely, suppose that $s_k = \sigma(k/n)$ with $\sigma(u)$ a positive, decreasing function on $[0,1]$ that satisfies

$$\lim_{u\to 0} u\sigma(u) = 0, \qquad \sup_{0\le u\le 1} |u\sigma'(u)| \le c_1.$$

Then, with $c_2 \le 1 + (c_1/\sigma(1))$,

$$\sum_{k=1}^{n} s_k \wedge \gamma_k \le c_2 \min_{0\le k\le n} \; [ks_k + \sum_{k+1}^{n} \gamma_n].$$

PROOF. Let $\Gamma_k = \sum_{i=k+1}^n \gamma_i$, and let $\kappa = \max\{k \ge 1 : s_k \le \gamma_k\}$ if such an index exists, otherwise set $\kappa = 0$. Using the monotonicity of both sequences and the definition of κ, we have

$$\begin{aligned} \sum_{i=1}^{n} s_i \wedge \gamma_i &= \sum_{i=1}^{\kappa} s_i \wedge \gamma_i + \Gamma_\kappa \ge \kappa(s_\kappa \wedge \gamma_\kappa) + \Gamma_\kappa \\ &= \kappa s_\kappa + \Gamma_\kappa \ge \inf_k \; ks_k + \Gamma_k. \end{aligned} \tag{13.19}$$

(b) For each k, including $k = 0$ and n, we have $\sum_{i=1}^n s_i \wedge \gamma_i \le \sum_{i=1}^k s_i + \sum_{i=k+1}^n \gamma_i$, so that

$$\sum_{i=1}^{n} s_i \wedge \gamma_i \le \min_{0\le k\le n} \sum_{i=1}^{k} s_i + \Gamma_k.$$

The result will follow if we show that $\sum_{i=1}^k s_i \le c_2 k s_k$ for $k = 0, \dots, n$. But

$$\sum_{i=1}^{k} s_i = \sum_{i=1}^{k} \sigma(i/n) \le \int_0^{k/n} \sigma(u)du.$$

Hence $(ks_k)^{-1} \le \sup_{0\le u\le}[x\sigma(x)]^{-1}\int_0^x \sigma(u)du$. The result now follows from

Remark. If $\sigma(u)$ is a positive, decreasing function on $[0,1]$ that satisfies

$$\lim_{u\to 0} u\sigma(u) = 0, \qquad \sup_{0\le u\le 1} |u\sigma'(u)| \le c_1.$$

Then, with $c_2 \le 1 + (c_1/\sigma(1))$ and $v \in [0,1]$,

$$\int_0^v \sigma(u)du \le c_2 v\sigma(v).$$

Indeed, by partial integration,

$$\begin{aligned}\int_0^v \sigma(u)du &= v\sigma(v) + \int_0^v u|\sigma'(u)|du \le v[\sigma(v) + c_1] \\ &\le v\sigma(v)[1 + c_1/\sigma(1)].\end{aligned}$$

□

13.7. False discovery rate estimation

The notion of *False Discovery Rate* (FDR) originated in the area of multiple inference, which is concerned with the simultaneous testing of a possibly large number of null hypotheses. When cast as a prescription for estimation, the FDR point of view leads to an estimator closely connected with the $2k\log(n/k)$ penalty class. We describe this connection, first by reviewing the simultaneous testing setting for FDR.

Consider the orthogonal regression model $y = \mu + \epsilon z$ in $\mathbb{R}^n$, and consider the n separate null hypotheses $H_i : \mu_i = 0$. Suppose that n independent test statistics are given, which we here take to be simply the components y_i of the data. Traditional *familywise error rate* (FWER) control methods seek to bound the chance of even one type I error among the n tests. The goal then is to guarantee that

$$P\{\text{reject} \ge 1 \ H_i \mid \text{all } H_i \text{ true}\} \le q,$$

for a specified value q. The standard way to achieve this is to use the Bonferroni approach: assign error probability q/n to each test and set the rejection regions as $\{|y_i| > t\}$ where t is chosen so that under the null distribution, $P\{|z_i| > t|\} = q/n$. As usual, let $z(\eta)$ denote the upper $100(1-\eta)\%$ quantile of the standard Gaussian distribution. Thus $t = t_B$, where

$$t_B = z(q/2n) \sim \sqrt{2\log(2n/q)} \sim \sqrt{2\log n}$$

for n large. This shows the fundamental limitation of the FWER control approach—since the thresholds are chosen quite high, the overall procedure has insufficient power to detect false H_i.

Here is an alternate sequential procedure, originally due to Seeger (1968) and Simes (1986), to decide which hypotheses to reject. Form the $2-$sided $P-$values $P_i = 2\tilde{\Phi}(|y_i|/\epsilon)$, order them: $P_{(1)} \le P_{(2)} \le \ldots \le P_{(n)}$, and look for the last crossing time of a linear boundary with slope q/n:

$$\hat{k}_F = \max\{i : P_{(i)} \le iq/n\}. \tag{13.20}$$

`INCLUDE DIAGRAM!` Now reject all null hypotheses $H_{(i)}$ corresponding to $i = 1, \ldots, \hat{k}_F$. By contrast, the Bonferroni method rejects some hypotheses if and only if $P_{(1)} \le$

q/n. If $\hat{k}_B = \#\{i : P_{(i)} \le q/n\}$ is the number of Bonferroni-rejected hypotheses, then necessarily $\hat{k}_F \ge \hat{k}_B$, so that FDR always generates at least as many "discoveries" as Bonferroni, and typically many more.

A major contribution of Benjamini & Hochberg (1995) was to provide a definition of false discovery rate that the $\hat{k}_F$ parameter satisfies. Let $\mathcal{N} = \{k : \mu_k = 0\}$ be the set of "true" null hypotheses, and $\hat{\mathcal{D}} = \{k : H_k \text{ is rejected}\}$, the set of "discoveries" made from the data. Then set

$$FDR = |\hat{\mathcal{D}} \cap \mathcal{N}|/|\hat{\mathcal{D}}|,$$

the fraction of discoveries that are false, i.e., come from true null hypotheses.

The key result of Benjamini & Hochberg (1995) is that the sequential procedure producing $\hat{k}_F$ guarantees that

$$E_\mu\{\ FDR\ \} \le q, \qquad \text{for all } \mu \in \mathbb{R}^n.$$

Thus, the expected fraction of spurious results is controlled, *regardless* of the configuration of the true means. [In fact, Benjamini & Hochberg (1995) prove a somewhat sharper result, with the upper bound being $n_0 q/n$, where $n_0 = \#\{i : \mu_i = 0\}$].

FDR Estimation. As proposed by Abramovich & Benjamini (1995), the definition (13.20) can be converted into a prescription for *estimation* in the sequence model via the switching relation

$$P_{(k)} = 2\Phi(|y|_{(k)}/\epsilon) \le kq/n \qquad \Leftrightarrow |y|_{(k)}/\epsilon > z(kq/2n). \tag{13.21}$$

(Here, $|y|_{(1)} \ge \ldots \ge |y|_{(n)}$ are the order statistics of the data y). This suggests that we define a boundary sequence $k \to t_{n,k}$ via the expression $t^2_{n,k} = z(kq/2n)$. This sequence is decreasing and satisfies

$$t^2_{n,k} \sim 2\log(2n/kq) \tag{13.22}$$

as $k/n \to 0$. From (13.20) and (13.21), we see that the FDR index $\hat{k}_F$ is the last crossing time

$$\hat{k}_F = \max\{k : |y|_{(k)} \ge \epsilon t_{n,k}\}. \tag{13.23}$$

The corresponding estimator is just hard thresholding with a *data determined* threshold $\hat{t}_F = t_{\hat{k}_F}$:

$$\hat{\mu}_{F,k} = \eta_H(y_k; \hat{t}_F) = \begin{cases} y_k & |y_k| \ge \hat{t}_F \\ 0 & \text{otherwise.} \end{cases}$$

Now the connection with penalized estimation is apparent: (13.23) is just the backward stepwise model selection method discussed in Section 13.5 for the threshold sequence $t^2_{n,k} = z(kq/2n)$. The associated penalized least squares criterion is given by (13.15), and the relation (13.22) indicates that the penalty may belong to the $2k\log(n/k)$ class. `FIX UP!`