Asymptotic selection of models for regression

Jean-Marc Azaïs, Université de Toulouse

February 24, 2012

1 Introduction

In many situations one uses several variables in a regression model as a precaution. Some of the k variables are relevant and some not. Fitting the whole model with say k variables seems to make sense but it often leads to disagreeable phenomenum : the over-fitting in the sense that the estimated coefficients follow the data and thus the errors. They don't follow the model. A good example is given by Figure 1 where a smooth signal is observed with some noise and is estimated by a piecewise constant model with 42 sub-intervals.

To avoid this kind of phenomenon, we must introduce the parsimony principle: to estimate too many parameters leads to an inflation of variance and a poor performance of the estimation of the response. Thus is is often better to set to zero the coefficients of some explanatory variables that seem to have a small or non-significative influence.

More precisely, we consider a regression model with k regressors and n observations.

$$Y_i = \sum_{j=1,..k} \beta_j Z_i^{(j)} + \epsilon_i, i = 1, ..., n$$

We will assume in most of the parts of this paper that $n > k$ and that the model is regular.

The whole model, the true model, the over-models and the false models

The model with all regressors will be called the "whole model". It will be denoted by $\bar{m}$. Among the coefficients $\beta_1, \ldots, \beta_k$ of the whole model, some may be zero and the corresponding variables are not needed. They will be called the superfluous variables. The goal of choice of model is to identify these variables and consequently the true model that consists of all variables with non-zero beta. This model will be denoted m^*; to identify m^* instead of $\bar{m}$ permits to avoid the over-fitting.

The identification can be false in two direction

- we can chose an "over-model": it is a model m that strictly contains m^*. As a consequence in contains some superfluous variables.
- we can chose a "false model" that does not contain some variables of m^*. For us a false model may contain some superfluous variables, it does not matter.

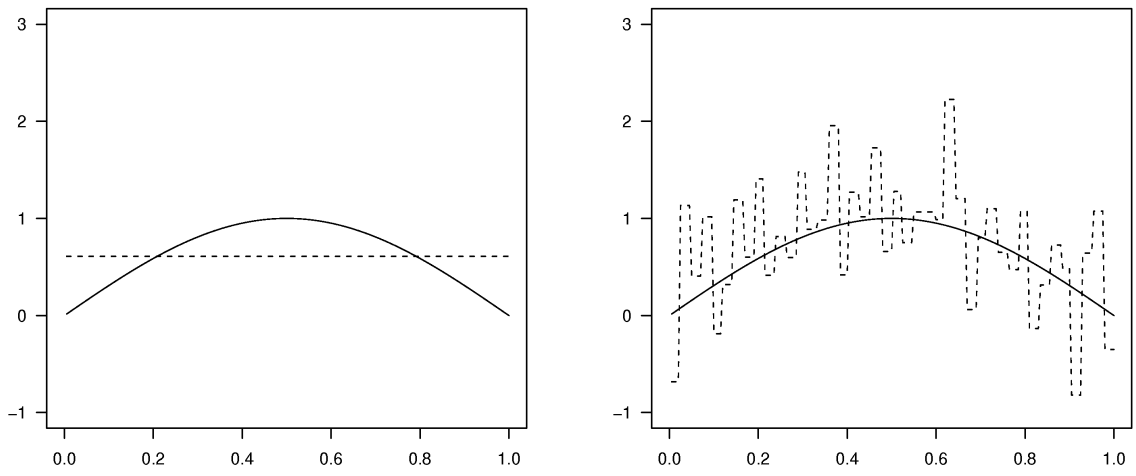

Figure 1: Over-fitting with a piecewise constant estimation over 42 sub-intervals.

As an example if $(\bar{m}) = \{1, 2, 3, 4, 5\}$ and $m^* = \{1, 4, 5\}$

- $\{1, 3, 4, 5\}$ is an over-model
- $\{1, 5\}$ and $\{1, 3, 5\}$are two false-models. The first one is a sub-model (of the true one) but it does not matter.

We will consider the "all sub-set regression" in the sense that will search the true model among the set $\mathcal{M}$ of the 2^k sub-models of $\bar{m}$. Some exception to that case are

- nested model, for example in polynomial regression : the jth regressor $Z_i^{(j)}, i = 1, ..., n$ is a power of the second regressor $Z_i^{(2)}$ (The first one is the constant) and we want to chose the degree of the polynomial. There are k sub-models only of $\bar{m}$ to consider.
- model with an intercept. In almost all the cases, the first regressor is the "all-one vector" $\mathbb{1}_n$ and in many case one does not want to question the presence of this vector in the model. In that case the set $\mathcal{M}$ of models to be considered is of size 2^{k-1}. This case is very similar to all sub-set regression, so we will omit the details.

Elementary methods

Test or thresholding.

Tests: Let two models m_1 and m_2 of the set $\mathcal{M}$ of considered models. Let α a level that may depend on n. When m_1 and m_2 are nested, one method is to perform a classical α F test between them. This method leads to two problems; first the number of tests to

perform is very large $(3^k - 2^k)$ and second it may be non-consistent in the sense that m_1 can be chosen preferable to m_2 and m_2 to m_3 while m_3 is chosen preferable to m_1

Thresholding: One very crude method is to adjust the whole model and perform an α T-test of each of the k variables and keep the significative ones. This certainly makes sense if the model (or equivently the regressors) is orthogonal. In the other case it can lead to strange decisions. For example if $Z^{(1)}$ and $Z^{(2)}$ are very collinear and very collinear to Y, the thresholding method will discard both variables because, when $Z^{(1)}$ is present, $Z^{(2)}$ is no longer needed and vice versa. Nevertheless we will be able to prove some properties of this method.

Backward regression: to avoid the problem encountered in the example above, the backward selection method starts with the whole model and then

- at each step, the least significant variable is removed from the model and calculations are made anew.

-this is done while the variable to be removed is non-significant at a α level. If the variable is significant, of course it is kept, the procedure stops and the model is chosen .

Stepwise regression is a variant of the preceding where at every step we may add or remove a variable. We skip the details.

Forward regression is exactly the contrary of backward regression: we start with the empty model or the model with the sole constant and we add at each step the most significative variable. We end when the variable to be add is non significative at α level. The forward regression which is also called "L^2 boosting " ref?? can be applied in the case $k > n$.

PRESS or cross-validation

Let $m \subset \bar{m}$ and let us consider the associated regression model

$$Y_i = \sum_{j \in m} \beta_j Z_i^{(j)} + \epsilon_i,$$

We denote by X_m the design matrix (the matrix of the linear model) associated to model $m \in \mathcal{M}$. We want to estimate the quadratic error of model m

$$\mathbb{E}\big(\|X_{m^*}\beta - X_m\widehat{\beta_m}\|^2\big).$$

A way of estimating this is a "leave-one-out cross-validation".

For $i = 1, \ldots, n$ we, define Y^{-i} the vector obtained be removing the ith observation. For $m \in \mathcal{M}$ we define μ_m^{-i} as the scalar which is the prediction

based on Y^{-i} and on Model m

and taken at the point i.

We define the PRESS as

$$\text{PRESS }_m = \sum_{i=1}^{n} (Y_i - \mu_m^{-i})^2. \tag{1}$$

Note that one nice property of the PRESS is that the two random variables Y_i and μ_m^{-i} are independent.

The chosen model is the one with the lowest PRESS.

In general, selection with PRESS is computationally very expensive and has properties equivalent to C_p or AIC (see above) see LI (1987) . In particular, as we will see, it tends to over-estimate the size of the model. For linear model the situation is nicer since we can compute a simplified form: it can be proved that

$$PRESS_m = \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i^m)^2}{(1-h_i)^2},$$

where $h_i = X_i'(X'X)^{-1}X_i$, X_i being the ith row of X see, for example, McQuarrie and Tsai(2007) p. 252 for a proof.

To remedy to this drawback one can use "leave-p-out cross-validation, but in this case the computational cost is even larger. A less costly alternative in v-fold cross-validation (Geisser 1975) , see Arlot (2009) for a detailled study .

2 Methods based on L_0 penalties

As a general principle the likelihood method choses always the largest model and this is true for our regression model. Note that, as we will see later maximizing the likelihood is equivalent, for linear model to minimize the sum of square. Heuristic considerations (based on Kullback information or on Bayesian models) have lead to use the following penalized likelihood criterions

$$AIC = -2\log(\text{maximized likelihood}) + 2|m|$$
$$BIC = -2\log(\text{maximized likelihood}) + \log(n)|m|$$

where

- the maximized likelihood is the maximum of the likelihood
- the likelihood is computed on n independent observations.
- The penalty $2|m|$ or $\log(n)|m|$ favors small models($|m|$ is the size of m).
- the criterions have to be minimized.

These criterion can be extended as

$$GIC = -2\log(\text{maximized likelihood}) + c(n)|m| \tag{2}$$

where $c(n)$ is a function to be fixed later.

We can define also

$$AIC_c = -2\log(\text{maximized likelihood}) + n\frac{n+|m|+1}{n-|m|-3}$$

and

$$C_p(m) = \frac{SS(m)}{\widehat{\sigma}_{\overline{m}}} + 2|m|$$

Lemma 1 *In the linear model*

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

we have

- *the maximum likelihood estimator of* σ^2 *is* $\widehat{\sigma^2} = 1/nSS$ *where* SS *is the sum of squares*

$$SS = \|Y - X\widehat{\beta}\|^2.$$

- $$-2\log(\text{ maximized likelihood}) = n\log(\widehat{\sigma^2}) + SS/\widehat{\sigma^2}$$

- $$-2\log(\text{ maximized likelihood}) = n\log(\widehat{\sigma^2}) + n$$

- $$-2\log(\text{ maximized likelihood}) = n\log(SS/n) + n = n\log(SS) + (const).$$

The proof is omitted, each result being an easy consequence of the preceding one. Note that the constants (*const*) appearing in the formulas above play no role and can be omitted.

All the criterions as PRESS AIC BIC GIC permit a rather easy comparison of models **but** if we perform a "all subset selection", the number of sub-model to compare is 2^k which is soon very large. Some "leaps and bounds algorithm" exist **??**** that permit to avoid to examine all the possibilities but practical limitations are about $k = 30$. In the other cases only a partial exploration is performed by a stepwise algorithm. This works rather well in practice but no theoretical results are known in that case. For these reasons for large size, one often prefers L^1 penalties as LASSO.

The criterion PRESS AIC BIC GIC permit to consider the "large dimension case" : $k > n$ only if we limit us to models of size $|m| < S$ with S much smaller than n . Such models are called sparse models. But in this case again computational problems are heavy.

3 Comparison of models with AIC, GIC

This section is devoted to the study of the relations of AIC (BIC, GIC) with tests.

Assumption 1 *Though we will use Gaussian likelihood to estimate and compute the criterion, we will work under one of the two following hypotheses when* n *tend to infinity and when* k *may depend on* n *but must satisfy* $k_n = o(n)$

- *the Gaussian case : the* $\epsilon_i, i = 1, \ldots, n$ *of the linear model errors are independent with distribution* $N(0, \sigma^2)$*, the variance* σ^2 *is of course unknown.*

- *the errors are centered independent with the same symmetric distribution and finite variance* σ^2 *and finite order four moment. We assume in addition the Huber condition :* H^n *the maximal diagonal element of the "hat matrix"* $X(X'X)^{-1}X'$ *tends to zero (*X *is the matrix associated to the whole model).*

3.1 AIC

Lemma 2 *Suppose that m_1 and m_2 are two nested models $m_1 \subset m_2$, the model m_1 is preferable to m_2 for AIC iff*

$$\widehat{F}_{m_2/m_1} = \frac{(SS(m_1) - SS(m_2))/(|m_2| - |m_1|)}{SS(m_2)/(n - |m_2|} > \frac{n - |m_2|}{|m_2| - |m_1|}\left(\exp\left(2\frac{(|m_2| - |m_1|)}{n}\right) - 1\right) \tag{3}$$

Again the proof can be omitted. The reader familiar with linear model has recognized in the left-hand-size of (3) the statistics of the Fisher test. In other words, AIC performs a Fisher test, but with a different critical value. We have obviously the same kind of result for GIC replacing the 2 by $c(n)$.

Suppose now that the number n tends to infinity that m_1 is the true or an over-model and that Assumption 1 is satisfied, then using law or large number and Central limit theorem under Lindeberg condition (see for example Th 8.3 of Azaïs and Bardet) (since we are under the null hypothesis) the limit distribution of $\widehat{F}_{m_2/m_1}$ is

$$\widehat{F}_{m_2/m_1} \Rightarrow \chi^2(p)/p$$

where $\Rightarrow$ is the convergence in distribution. Obviously the right-hand side of (3) converges to 2.

As a consequence AIC performs asymptotically a F test with critical value $2p$ where p is the difference of degrees of freedom between the two hypotheses.
This corresponds to the following levels.

difference	level
1	0.104
2	0.068
3	0.049
4	0.037
5	0.028

As a consequence, considering the case where m_1 is the true model, the result above shows that AIC has a probability that tends to a positive limit to prefer every over-model m_2 to the true model. So the probability of choosing m^* cannot tend to 1.

3.2 BIC, GIC

If we consider the GIC as defined by (2). The calculation above shows that the criterion will prefer model m_1 to m_2 iff

$$\widehat{F}_{m_2/m_1} = \frac{(SS(m_1) - SS(m_2))/(|m_2| - |m_1|)}{SS(m_2)/(n - |m_2|)} > \frac{n - |m_2|}{|m_2| - |m_1|}\left(\exp(c(n)\frac{(|m_2| - |m_1|)}{n}) - 1\right)$$

Now we assume that $c(n) \to +\infty$, $c(n) = o(n)$ to get that the right hand side is equivalent to $c(n)$.

As a consequence the probability of preferring a given over-model is, for n suffiently large, smaller that the probability of a $\xi^2(d)$ distribution to be smaller that K for every K so it tends to zero.

I Sine k is assumed to be fixed, the number of over -models is bounded and we obtain immediately that the probability of "preferring an over-model to m^*" tends to zero.

3.3 Case of a false model

Suppose that m is "false". It is not in general a sub-model of the true model. But it can be compared to the model $m_2 = m \cup m*$. Since $m \subset m_2$ we can appy Lemma 2 that shows that asymptotically m_2 is prefered to m if $\hat{F}_{m_2/m} \geq \widetilde{c}(n)$ where $\widetilde{c}(n) \simeq c(n)$. Using the same arguments of chapter 8 of Azaïs and Bardet (2005), we see that under our hypotheses the denominator $D = \widehat{\sigma}^2$ of $\hat{F}_{m_2,m}$ tends in probability to σ^2 while the numerator can be written as

$$N = \left(\|P_V(X\beta) + P_V X\widehat{\beta}\|^2\right)/d$$

where V is the orthogonal of $[X_m]$ in $[X_{m_2}]$. Using the normality of $\widehat{\beta}$ (Th 8.2 of the same book) we see that what ever the non-centrality parameter $\|P_V(X\beta)$ is, the numerator can be written as

$$D = 1/d\|\|P_V(X\beta) + Z\|^2$$

where Z has for variance-covariance matrix

$$P_V X(X'X)^{-1}X'P_V = P_V.$$

Using a rotation argument, this matrix can be transformed, for example, into I_d where I_d is the indentity of size d and N can be written as the norm a vector in a space of dimension d as

$$N = \frac{\sigma^2}{d}\|\xi + W_n\|^2$$

where ξ converges to the $N(0, I_d)$ distribution and

$$\|W_n\|^2 = \frac{1}{\sigma^2}\|P_V X\beta\|^2 = \frac{1}{\sigma^2}\|P_{m^\perp} X\beta\|^2.$$

This last parameter will be called the no- centrality parameter and denoted by NC_m:

$$NC_m = \frac{\|P_{m^\perp} X\beta\|^2}{\sigma^2}$$

Note that this parameter depends additionally on n but we omit that in the notation.

We set now our uniform born on χ^2 distributions

Lemma 3 - *For all integer $d \geq 1$ and for all real $c(n)$ greater than 2*

$$\mathbb{P}\{\chi^2(d) \geq c(n)d\} \leq \exp(-c(n)/2)$$

- *If $NC > 4c(n)d$ then*

$$\mathbb{P}\{\chi'^2(d, NC) \leq c(n)d\} \leq \exp -(\frac{NC}{8d})$$

Proof: The first part is easy to obtain by an exponential inequality or by exact computation using integration by parts. Let χ'^2 be a variable with distribution $\chi'^2(d, NC)$. This variable has the representation

$$\chi'^2 = \|\sqrt{NC}e_1 + Z\|^2,$$

where e_1 is the first vector of the basis and Z is standard normal in $\mathbb{R}^d$. Let $\chi^2 = \|Z\|^2$. Denoting $c(n)$ by c for short, we have

$$\mathbb{P}\{\chi'^2 < cd\} = \mathbb{P}\{\chi' < \sqrt{cd}\} \leq \mathbb{P}\{\chi > \sqrt{NC} - \sqrt{cd}\} \leq \mathbb{P}\{\chi > 1/2\sqrt{NC}\} = \mathbb{P}\{\chi^2 > 1/4NC\}.$$

It suffices to use the first relation. ∎

We turn now to the main results. Suppose that the parameter $c(n)$ satisfies $1 << c(n) << n$ and that every false model m has a non-centrality parameter that satifies $c(n) << NC_m$, then

(i) Suppose now that m is a false model then using the convergence in probability to σ^2 of the denominator of $\hat{F}_{m_2,m}$ we obtain that

$$\begin{aligned}\mathbb{P}\{m \text{ prefered to } m_2\} &= \mathbb{P}\{\|\xi + W_n\|^2 \leq dC(n)(1 + o_p(1))\} \\ &\leq \mathbb{P}\{\|\xi\| \geq d\big(\sqrt{NC_m} - \sqrt{c(n)(1 + o_p(1))}\big)\} \\ &= \mathbb{P}\{\|\xi\| \geq d\big(\sqrt{NC_m}(1 + o_p(1))\big)\},\end{aligned}$$

where W and ξ are defined as above. The convergence in distribution of ξ implies that this probability tends to zero.

As a consequence mixing this case with the case of over-models, we have proven that GIC chooses the true model with a probability that converges to 1.

(ii) Suppose now in addition that the model is Gaussian,then it is possible to give an exponential bounds to the probability of a false model m to be preferred to $m_2 = m \cup m^*$. The false model m is prefered to $m_2 = m \cup m^*$ if $\hat{F}_{m_2,m} \leq \widetilde{C}_n$. Firstly We can choose n sufficiently large so that $\widetilde{C}_n \leq 2C(n)$, secondly we have

$$\hat{F}_{m_2,m} \stackrel{D}{=} \frac{\chi'^2(d, NC_m)/d}{\chi^2(n - |m_2|)/(n - |m_2|)}$$

Using large deviation inequality (in fact just the easier part) , except with an exponentially small (as a function of n) probability,

$$\chi^2(n - |m_2|) \leq (2(n - |m_2|))$$

so that it suffiices to give bound to

$$\mathbb{P}\{\chi'^2(d, NC_m)/d \leq 4C(n)\}$$

and this by lemma 3 is smaller than $\exp\big(-\frac{NC_m}{8d}\big)$ so we have proved that

$$\mathbb{P}\{m \text{ prefered to } m_2\} \leq \exp(-(constn) + \exp\big(-\frac{NC_m}{8d}\big)$$

A first example of application is the very simple case where k is fixed and the matrix X associated to the whole model satisfies

$$1/n\ X'X \to M \tag{4}$$

where M is some definite positive matrix. In that case the computation below proves that for every false model m

$$NC_m \simeq \gamma_m n$$

with $\gamma_m > 0$.

Computation of NC

Indeed by the Pythagore Theorem

$$NC_m = \|P_{m^\perp} X\beta\|^2 = \|X\beta\|^2 - \|P_m X\beta\|^2$$

We study the two terms separately. Because of our hypothesis

$$\|X\beta\|^2 \simeq n\beta' M\beta.$$

For the seond term

$$\|P_m X\beta\|^2 = \beta' X' X_m (X'_m X_m)^{-1} X'_m X\beta \simeq n\beta' M_{\overline{m},m} M^{-1}_{m,m} M_{m,\overline{m}}\beta$$

where M_{m_1,m_2} is the extraction of the matrix M choosing m_1 for the lines and m_2 for the columns. So that

$$\|P_{m^\perp} X\beta\|^2 \simeq n\beta'(M - M_{\overline{m},m} M^{-1}_{m,m} M_{m,\overline{m}})\beta$$

M can bee seen as the Gram matrix (the matrix of norms and scalar products) of some set of k vectors in $\mathbb{R}^k$, say $V_1, \ldots, V_k$ (the choice is up to a rotation). Since M is non singular these vectors are not collinear. A classical linear algebra calculation shows that

$$M - M_{\overline{m},m} M^{-1}_{m,m} M_{m,\overline{m}}$$

is the matrix of the quadratic form that associate to the vector $b \in \mathbb{R}^k$ the quantity

$$\|\Pi_{m^\perp} \sum_{i=1}^{k} b_i V_i\|^2,$$

where $\Pi_{m^\perp}$ is the projector on the orthogonal of the space S_m generated by the vector that are in m. Since m is a false model β has some coordinates that does not belong to m and because of the linear independence of the vectors, $\sum \beta_i V_i$ does not belong to S_m. and we obtain the result. ∎

Note that

- the condition (4) is met for example if the regressors are draw from i.i.d. replicates of some random distribution with a second order moment and non-degenerate variance matrix. This is a direct consequence of the law of large numbers.

- under this condition, it is an exercise, to check that the thresholding method and the backward method find the true model with a probability that tends to 1, as soon as the tests are conducted at a level α_n that tends to zero sufficiently slowly.
- The result can be generalized to the case where some normalization $d(n)$ of the information matrix exists such that

$$1/d(n)\ X'X \to M$$

  In such a case $c(n)$ must be negligible with respect to $d(n)$.
- When $k = k_n$ tends to infinity, we cannot hope to have a property like (4) but our result still prove that if $k(n) = o(c(n))$, GIC will chose and over-model with a probability that tends to zero.

4 Asymptotic oracle inequality

In this section we assume normality. Consider the quadratic risk of estimation and we still assume a) condition (4) with M regular b) that k fixed and c) that $1 << c(n) << n$. Let $\widehat{m}$ the model chosen by GIC (with probability 1 it is unique). We define the risk of estimation

$$R_n = \mathbb{E}\big(\|\widehat{Y}_{\widehat{m}} - X\beta\|^2\big).$$

This risk can be partitionned into the risks relative to the choice of a particular model

$$R_n = \sum_{m\in\mathcal{M}} R_n(m) := \sum_{m\in\mathcal{M}} \mathbb{E}\big(\|\widehat{Y}_m - X\beta\|^2\ \mathbb{1}_{\widehat{m}=m}\big)$$

Using the decomposition bias, variance, a short computation shows that
if Z is some random variable that can be written

$$Z = \mathbb{E}(Z) + \epsilon = \mu + \epsilon$$

with ϵ symmetric and E an event that may depend on ϵ but whose distribution is invariant by change of sign of ϵ, then

$$\mathbb{E}(Z\ \mathbb{1}_E)^2 = \mu^2\mathbb{P}(E) + \mathrm{Var}(\epsilon\ \mathbb{1}_E) + 2\mu\mathbb{E}(\epsilon\ \mathbb{1}_E) = \mu^2\mathbb{P}(E) + \mathrm{Var}(\epsilon\ \mathbb{1}_E).$$

Remarking that a change of sign of the errors does not modify the choice of model and using the assumed symmetry of the errors we get

$$\mathbb{E}\big(\|\widehat{Y}_m - X\beta\|^2\ \mathbb{1}_{\widehat{m}=m}\big) = \|P_{m^\perp}(X\beta)\|^2\mathbb{P}(\widehat{m}=m) + \mathbb{E}\big(\|P_m\epsilon\|^2\ \mathbb{1}_{\widehat{m}=m}\big) = J_{1,m} + J_{2,m}$$

Then every false model m satisfies

$$\|P_{m^\perp}(X\beta)\|^2 \simeq \gamma_m n \qquad \text{with } \gamma_m > 0.$$

Thus by Lemma 3 , for n sufficiently large$NC_m \simeq \frac{\gamma_m n}{\sigma^2} > 4c(n)$:

$$\mathbb{P}(\widehat{m} = m) \leq \exp -(\frac{\gamma_m n}{8 d_m}),$$

Where d_m is the number of missing variables in m: $d_m = |m \cup m^*| - |m|$. For over-models the quantity $\|P_{m\perp}(X\beta)\|^2$ vanishes. This implies that

$$\sum_{m \in \mathcal{M}} J_{1,m} \to 0.$$

For the other terms we use the Schwarz inequality

$$\mathbb{E}\big(\|P_m \epsilon\|^2 \ \mathbb{1}_{\widehat{m}=m}\big) \leq \big(\mathbb{E}\|P_m\epsilon\|^4 \mathbb{P}(m \neq m^*)\big)^{1/2}.$$

Let us compute the quantity $\mathbb{E}\|P_m\epsilon\|^4$. Let P_{ij} denote the entry i, j of P_m

$$\begin{aligned}\mathbb{E}\|P_m\epsilon\|^4 &= \sum_{iji'j'} \mathbb{E}(\epsilon_i P_{ij}\epsilon_j \epsilon_{i'} P_{i'j'}\epsilon_{j'}) \\ &= \sum_{iji'j'} P_{ij}P_{i'j'}\mathbb{E}(\epsilon_i\epsilon_j\epsilon_{i'}\epsilon_{j'}).\end{aligned}$$

Because of independance, the last expectation vanishes except if the four indices are pairwise equals. It remains three cases to consider

- $i = i' = j = j'$ which contribution is $m_4 \sum_i P_{ii}^2$ where m_4 is the order 4 moment of the errors.
- $i = j \neq i' = j'$ which contribution is $\sigma^4 \sum_{i \neq i'} P_{ii} P P_{i'i'}$
- $i = i' \neq j = j'$ or $i = j' \neq j = i'$ which contribution is bounded by $2\sigma^4 \sum_{i \neq j} P_{ij}^2$.

Since

$$\sum_i P_{ii}^2 + \sum_{i \neq j} P_{ij}^2 = tr(P_m^2) = tr(P_m) = |m|$$

$$\sum_{i \neq i'} P_{ii} P P_{i'i'} + \sum_i P_{ii}^2 = (tr(P_m))^2 = |m|^2,$$

it is easy to see that $\mathbb{E}\|P_m\epsilon\|^4$ is bounded. Note that in the Gaussian case it

is the expectation of square of a $\chi^2(|m|)$ variable which can be easily computed to be $\sigma^4(|m|^2 + 2|m|)$. Finally

$$\sum_{m \in \mathcal{M}, m \neq m^*} J_{2,m} \to 0`$$

Finally we have proven that

$$R_n = \mathbb{E}\big(\|\widehat{Y}_{\widehat{m}} - X\beta\|^2\big) \to |m^*|$$

The risk we have if we know the true model. The risk with a choice of model by GIC is asymptotically the same than the risk when the oracle tell us which is the true model. Such an inequality is called an Oracle inequality

Bibliography

Arlot S. Rééchantillonnage et Sélection de modèles. Phd these. Université de Paris -Sud.

Azaïs, J-M. and Bardet J-M. (2005). *Le modèle linéaire par l'exemple*. Dunod, France

Li, Ker-Chau. Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, Vol. 15, No. 3. (Sep., 1987), pp. 958-975.

McQuarrie A. and Tsai C-L regressio and time series model selection. World scientific

Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, 6(2); 461-464.