

Choix de modèle : une approche bayésienne pour un faible nombre de données

Guy DEMOMENT

Laboratoire des signaux et systèmes, UMR 8506 (CNRS/SUPÉLEC/UPS)
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France
demoment@lss.supelec.fr

Résumé — Le choix de l'ordre d'un modèle peut être fait en calculant un facteur de Bayes. Le critère de Schwarz en est une approximation utile, mais grossière. Un calcul plus précis peut être sensible au choix de l'a priori sur les paramètres des différents modèles en compétition. Un nouveau critère est proposé pour les problèmes de régression linéaire, fondé sur une approximation de Laplace. Son intérêt est établi quand le nombre de paramètres est une fraction importante du nombre de données.

Abstract — The assessment of evidence in favour of a model can be accomplished by calculating a Bayes factor. The Schwarz criterion remains a useful yet crude approximation. More accurate methods may be sensitive to the choice of priors on parameters appearing in the competing models. A new criterion is derived in linear regression problems, based on Laplace's approximation. It is of particular use when the number of parameters is a moderate to large fraction of the sample size.

1 Introduction

Le choix d'un modèle est une étape fondamentale de l'analyse de données expérimentales en général, et du traitement du signal et de l'image en particulier. Il a engendré une énorme littérature, avec beaucoup de solutions différentes [8]. Il bénéficie d'un regain d'intérêt à chaque fois que de nouvelles familles de modèles apparaissent en traitement du signal ou de l'image. Le but de cette communication est de rappeler les défauts des critères d'information les plus usités et de montrer l'intérêt des facteurs de Bayes dans les problèmes caractérisés par un faible nombre de données.

2 Vraisemblance pénalisée

Supposons que les données aient été engendrées par le « vrai » modèle : $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, avec $\mathbf{y} = [y_1, \dots, y_n]^t$, $\boldsymbol{\mu} = [\mu_1, \dots, \mu_n]^t$, $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_n]^t$, et où les ϵ_i sont i.i.d. et $\sim \mathcal{N}(0, \sigma_0^2)$. Considérons une famille de modèles candidats

$$M_p : \mathbf{y} = \mathbf{h}_p(\boldsymbol{\theta}_p) + \mathbf{e}, \quad (1)$$

avec $\mathbf{e} = [e_1, \dots, e_n]^t$, $\mathbf{h}_p(\boldsymbol{\theta}_p) = [h_{1p}(\boldsymbol{\theta}_p), \dots, h_{np}(\boldsymbol{\theta}_p)]^t$, où $\boldsymbol{\theta}_p$ est un vecteur de dimension p et où les e_i sont i.i.d. et $\sim \mathcal{N}(0, \sigma^2)$.

Dans la littérature classique sur le sujet, le critère primordial de choix est le rapport de vraisemblance. L'approche classique n'a pas grand chose à dire sur la comparaison des modèles non emboîtés, mais dans le cas des modèles emboîtés, le rapport de vraisemblance n'est pas une statistique convergente, il tend à donner trop de poids au modèle d'ordre le plus élevé. En conséquence de nombreux auteurs ont proposé de pénaliser la vraisemblance maximisée avec des fonctions croissantes de l'ordre.

Une mesure d'écart entre modèle vrai et modèle candidat souvent invoquée est constituée par la distance de Kullback entre les distributions pour les données sous l'une ou l'autre des deux hypothèses, ou distributions *directes*. Cette distance n'étant pas calculable en pratique, une des méthodes de choix les plus utilisées en traitement du signal — le critère d'information d'Akaike (AIC) [1] — l'estime moyennant certaines hypothèses supplémentaires. Il améliore le rapport de vraisemblance mais il conduit systématiquement à une probabilité de surparamétrisation strictement positive. Un autre critère très populaire, introduit par Rissanen sur des bases très différentes, le principe de *longueur de description minimale* (MDL) [11], et identique à celui de Schwarz (BIC) [12] fondé sur une approche bayésienne, pénalise la log-vraisemblance à son maximum par une fonction linéaire de l'ordre du modèle, et apporte une amélioration par une estimation convergeant presque sûrement. Il en existe bien d'autres, mais la forme générale de ces critères d'information peut se résumer à :

$$IC(p) = -2L(\hat{\boldsymbol{\theta}}_p) + \alpha(n)p \quad (2)$$

où $L(\hat{\boldsymbol{\theta}}_p)$ désigne la log-vraisemblance maximisée du modèle M_p . Les formes possibles de pénalisation sont $\alpha(n) = 2$ pour AIC, $\alpha(n) = \log n$ pour BIC, $\alpha(n) = \log \log n$ pour le critère de Hannan & Quinn, etc ...

Bien que certains de ces critères possèdent des propriétés d'optimalité asymptotique [13], ils souffrent de défauts qui deviennent particulièrement évidents quand le nombre n de données est faible : ils tendent à surestimer considérablement l'ordre à moins que de fortes restrictions ne soient mises à la dimension maximale des modèles candidats. De plus, les résultats obtenus avec un faible nombre de données sont très sensibles à la forme de la pénalisation apportée à la log-vraisemblance [5]. Mais la difficulté du problème fait que les études à nombre fini de données sont très rares [5] [2].

3 Approche bayésienne

Dans cette approche [6, 14, 7], les modèles statistiques sont introduits pour représenter la probabilité des données en fonction de chacun des modèles (probabilités directes), et le théorème de Bayes est employé pour calculer la probabilité à postériori que l'un des modèles soit correct. L'élément central est le *facteur de Bayes* dont le calcul, comme dans toute méthode bayésienne, implique des intégrales et donc, souvent, des intégrations numériques.

3.1 Facteurs de Bayes

Supposons pour commencer que les données \mathbf{y} aient été engendrées sous l'une ou l'autre de deux hypothèses M_1 et M_2 , selon une densité de probabilité $f(\mathbf{y} | M_1, I)$ ou $f(\mathbf{y} | M_2, I)$. Étant donné des probabilités à priori $P(M_1 | I)$ et $P(M_2 | I) = 1 - P(M_1 | I)$, les données produisent les probabilités à postériori $P(M_1 | \mathbf{y}, I)$ et $P(M_2 | \mathbf{y}, I) = 1 - P(M_1 | \mathbf{y}, I)$. L'évidence apportée par les données en faveur de l'une ou l'autre hypothèse prend une forme particulièrement simple en utilisant l'échelle des *cotes* (cote = probabilité / (1 - probabilité)) :

$$\frac{P(M_1 | \mathbf{y}, I)}{P(M_2 | \mathbf{y}, I)} = \frac{P(M_1 | I) f(\mathbf{y} | M_1, I)}{P(M_2 | I) f(\mathbf{y} | M_2, I)}. \quad (3)$$

La cote à priori est donc simplement multipliée par

$$B_{12} = \frac{f(\mathbf{y} | M_1, I)}{f(\mathbf{y} | M_2, I)}, \quad (4)$$

qui est le *facteurs de Bayes* [7]. Dans le cas le plus simple, quand les deux hypothèses sont des distributions sans paramètre libre (hypothèses *simples*), B_{12} est le rapport de vraisemblance. Sinon, quand l'une ou l'autre des hypothèses implique des paramètres inconnus, le facteur de Bayes a toujours la forme d'un rapport de vraisemblance, mais les densités $f(\mathbf{y} | M_k, \theta_k, I)$ ($k = 1, 2$) sont obtenues en *intégrant* (et non en maximisant) sur l'espace des paramètres :

$$f(\mathbf{y} | M_k, I) = \int f(\mathbf{y} | M_k, \theta_k, I) f(\theta_k | M_k, I) d\theta_k. \quad (5)$$

Cette vraisemblance *marginale* n'est rien d'autre que le coefficient de normalisation dans l'expression de la densité de probabilité à postériori pour le paramètre θ_k donnée par la règle de Bayes, mais contrairement à beaucoup d'autres applications de la vraisemblance, tous les facteurs constants intervenant dans son expression doivent être conservés lors du calcul de B_{12} .

S'il y a maintenant P modèles M_1, \dots, M_P , et si l'on définit l'hypothèse *nulle* $M_0 : \mathbf{y} = \mathbf{e}$, la relation (3) montre que la probabilité à postériori pour M_p est donnée par

$$P(M_p | \mathbf{y}, I) = \frac{w_p B_{p0}}{\sum_{k=0}^P w_k B_{k0}} \quad (6)$$

où $w_k = P(M_k | I) / P(M_0 | I)$. Quand les différentes hypothèses M_p sont également probables à priori, les probabilités à postériori sont proportionnelles aux facteurs de Bayes, avec pour toutes le même coefficient. Un passage au logarithme rend les relations (3,4) additives et montre le rôle essentiel de la log-vraisemblance marginale :

$$EV(p) = \log f(\mathbf{y} | M_p, I) \quad (7)$$

ou *évidence* pour le modèle candidat M_p .

On peut montrer [7, 10] que le critère de Schwarz est une approximation grossière de facteurs de Bayes correspondant à des à priori variés. Il offre l'avantage appréciable d'éviter d'avoir à préciser l'à priori, au prix des inconvénients évoqués plus haut. Mais tout affinage soulève deux difficultés.

3.2 Intégration

Les distributions à priori $f(\theta_k | M_k, I)$ sont techniquement nécessaires. Ce peut être considéré comme un bien ou un mal. Un bien car cela permet, comme pour l'estimation de paramètre à coût minimal, de lever un certain nombre de difficultés propres à l'approche classique, et d'inclure d'autres informations sur les valeurs des paramètres. Un mal car ces distributions à priori peuvent être délicates à choisir. Une seconde difficulté dans l'utilisation des facteurs de Bayes est évidemment dans le calcul analytique de l'évidence. Un calcul exact est possible dans le cas des modèles linéaires gaussiens pour des distributions de la famille exponentielle, avec un à priori conjugué.

On peut aussi, comme [9, 3], employer une expression approchée de l'évidence, fondée sur la méthode de Laplace qui consiste à effectuer un développement de Taylor au second ordre du logarithme de l'intégrande autour de son mode, puis à en prendre l'exponentielle avant d'intégrer. Lorsque la distribution à priori $f(\theta_p | M_p, I)$ ne varie pas de manière appréciable sur le domaine où la vraisemblance est piquée, une variante usitée est :

$$f(\mathbf{y} | M_p, I) \approx f(\mathbf{y} | M_p, \hat{\theta}_p, I) \dots \times f(\hat{\theta}_p | M_p, I) (2\pi)^{p/2} |\mathbf{I}_e|^{-1/2} \quad (8)$$

où $\hat{\theta}_p$ est l'estimateur du maximum de vraisemblance et où \mathbf{I}_e est la matrice d'information empirique.

On peut enfin utiliser des méthodes de Monte Carlo, simples ou améliorées par échantillonnage d'importance, mais elles sont plus lourdes à mettre en œuvre que les méthodes d'approximation analytique, et pas nécessairement plus précises dans le cas simple mais très répandu d'un modèle linéaire [4].

3.3 Choix de l'à priori

Lors du choix et de la manipulation des distributions à priori, des simplifications peuvent être faites, mais elles doivent l'être avec soin. Pour l'estimation de paramètre, l'à priori est souvent choisi par commodité car on sait que si les données sont assez nombreuses, alors l'effet de l'à priori est faible. Pour le test d'hypothèse, il n'en va pas de même. La principale difficulté provient de l'emploi d'à priori diffus impropres. En effet, contrairement au cas du calcul de la distribution à postériori pour le paramètre $f(\theta_p | \mathbf{y}, M_p, I)$ qui reste propre lorsqu'on utilise un à priori impropre avec une vraisemblance dont la décroissance est assez rapide à l'infini, les à priori impropres introduisent un problème de mise à l'échelle de l'évidence, conséquence du changement de la dimension du paramètre d'un modèle à l'autre. L'évidence est définie à un facteur arbitraire près, et le facteur de Bayes obtenu est arbitraire

[3]. Un moyen de contourner cette difficulté est d'utiliser une partie des données pour un *apprentissage*, avec un à priori impropre, pour obtenir une distribution à postériori propre qui sera choisie comme à priori pour calculer le facteur de Bayes avec les données restantes [7].

4 Nouveau critère

On se limite au cas de la régression linéaire : $\mathbf{h}_p(\boldsymbol{\theta}_p) = \mathbf{X}_p \boldsymbol{\theta}_p$ qui recouvre les situations pratiques les plus fréquentes. La vraisemblance est donc (en abandonnant les indices quand il n'y a pas d'ambiguïté) :

$$f(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}, \sigma, M_p) \propto \sigma^{-n} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}{2\sigma^2} \right].$$

Cette vraisemblance est maximale pour $\hat{\boldsymbol{\theta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$ (nous supposons donc que la matrice normale $\mathbf{X}^t \mathbf{X}$ est inversible) et $\hat{\sigma}^2 = S/n$, où S est la somme des carrés des erreurs résiduelles. La log-vraisemblance prend donc, à son maximum, la valeur $L(\hat{\boldsymbol{\theta}}) = -(n/2) \log S + \text{const}$.

4.1 À priori de Cauchy

Il est souhaitable d'employer un à priori le moins informatif possible et qui dépende seulement d'hypothèses structurelles. Il existe quelques exigences minimales pour son emploi dans des facteurs de Bayes [6]. La première est qu'il soit propre, ainsi que nous l'avons signalé plus haut. Pour éviter que le test ne soit biaisé en faveur de l'alternative, l'à priori doit être localisé près de la valeur nulle $\boldsymbol{\theta} = 0$. Finalement, s'il possède une covariance, celle-ci doit représenter une information vague qui doit être mise à l'échelle par, mais inférieure à, l'information sur $\boldsymbol{\theta}$ contenue dans la vraisemblance. Un choix possible est la distribution normale conjuguée dont la covariance est l'inverse de l'information de Fisher, mais il conduit au paradoxe que le facteur de Bayes ne peut être inférieur à une fonction positive de n , quel que soit le bon accord des données entre elles. Nous suivrons donc plutôt la recommandation de Jeffreys qui est d'employer une distribution de Cauchy multi-variée, la plus simple satisfaisant des exigences spécifiques aux facteurs de Bayes [6]. Nous choisissons ainsi :

$$f(\boldsymbol{\theta}, \sigma | M_p, I) = f(\boldsymbol{\theta} | \sigma, M_p, I) f(\sigma | I) \quad (9)$$

où $f(\boldsymbol{\theta} | \sigma, M_p, I)$ est la distribution de Cauchy localisée à l'origine et de paramètre d'échelle $\mathbf{X}^t \mathbf{X}/n\sigma^2$ car rien, à part σ , ne peut fixer d'échelle pour $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta} | \sigma, M_p, I) = \mathcal{C}_{\text{au}}(0, \mathbf{X}^t \mathbf{X}/n\sigma^2) = A B(\sigma) C(\boldsymbol{\theta}, \sigma)$$

avec $A = \Gamma(p/2)/\pi^{p/2}$, $B(\sigma) = |\mathbf{X}^t \mathbf{X}/n\sigma^2|^{1/2}$ et $C(\boldsymbol{\theta}, \sigma) = [1 + \boldsymbol{\theta}^t \mathbf{X}^t \mathbf{X} \boldsymbol{\theta}/n\sigma^2]^{-p/2}$, et où $f(\sigma | I)$ est la distribution de Jeffreys :

$$f(\sigma | I) \propto \sigma^{-1}.$$

Cette dernière distribution est impropre, mais comme elle concerne le même paramètre de nuisance σ pour tous les modèles candidats, ceci est sans conséquence.

4.2 Intégration

Le calcul de l'évidence pour le modèle M_p nécessite celui de l'intégrale

$$\begin{aligned} I &= \int f(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}, \sigma, M_p, I) f(\boldsymbol{\theta}, \sigma | M_p, I) d\boldsymbol{\theta} d\sigma \\ &= \int_0^\infty A B(\sigma) \sigma^{-(n+2)} \exp \left[-\frac{\mathbf{y}^t \mathbf{y} - \hat{\boldsymbol{\theta}}^t \mathbf{X}^t \mathbf{X} \hat{\boldsymbol{\theta}}}{2\sigma^2} \right] d\sigma \\ &\dots \int_{\mathbb{R}^p} C(\boldsymbol{\theta}, \sigma) \exp \left[-\frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \mathbf{X}^t \mathbf{X} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{2\sigma^2} \right] d\boldsymbol{\theta}. \end{aligned}$$

La méthode de Laplace n'est appliquée qu'à l'intégration par rapport à $\boldsymbol{\theta}$. Les grandes valeurs du second intégrande, à σ fixé, sont dans le volume $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} \pm O(\sigma[\mathbf{X}^t \mathbf{X}]^{-1/2} \mathbf{u})$ où $C(\boldsymbol{\theta}, \sigma)$ varie peu par rapport à sa valeur en $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. D'où la première approximation :

$$I \approx \int_0^\infty A \left[\frac{2\pi}{n} \right]^{p/2} \sigma^{-n-2} \exp(-S/2\sigma^2) C(\hat{\boldsymbol{\theta}}, \sigma) d\sigma.$$

Le terme en $\hat{\boldsymbol{\theta}}^t \mathbf{X}^t \mathbf{X} \hat{\boldsymbol{\theta}}/n\sigma^2$ dans $C(\hat{\boldsymbol{\theta}}, \sigma)$ variant lentement avec σ , on y remplace $n\sigma^2$ par S , d'où

$$I \approx A \left[\frac{2\pi}{n} \right]^{p/2} C(\hat{\boldsymbol{\theta}}, \hat{\sigma}) \int_0^\infty \sigma^{-n-2} \exp(-S/2\sigma^2) d\sigma.$$

L'intégrale restante conduit classiquement à une fonction gamma, si bien que

$$I \approx \Gamma(p/2) n^{-p/2} C(\hat{\boldsymbol{\theta}}, \hat{\sigma}) (S/2)^{-(n+1)/2} \Gamma((n-3)/2),$$

sous réserve que $n > 8$. En ne gardant que les termes dépendant de p et en passant au logarithme népérien, on obtient finalement une expression approchée de l'évidence :

$$\begin{aligned} EV(p) &\approx -\frac{n+1}{2} \log S + \log \Gamma \left(\frac{p}{2} \right) \dots \quad (10) \\ &\quad -\frac{p}{2} \left[\log \left(\frac{n}{2} \right) + \log \left(1 + \frac{\hat{\boldsymbol{\theta}}^t \mathbf{X}^t \mathbf{X} \hat{\boldsymbol{\theta}}}{S} \right) \right]. \end{aligned}$$

On retrouve le critère BIC, mais avec des termes de pénalisation supplémentaires dont l'expression dépend, entre autres, de la structure du modèle au travers de la matrice d'information empirique.

4.3 Convergences

La question de la convergence de ce critère se pose, bien entendu, mais la réponse est double, comme dans toute approche bayésienne :

— On peut montrer, en utilisant le théorème de Stirling et un développement de Taylor de $\log(1+x)$, que l'erreur d'approximation de l'évidence, et donc du facteur de Bayes, par (10) est en $O(n^{-1/2})$, comme pour d'autres approximations, y compris celle du critère BIC.

— L'estimateur est convergent au sens classique dans la mesure où, quand $n \rightarrow \infty$, $\log B_{12} \rightarrow +\infty$ sous M_1 , alors que $\log B_{12} \rightarrow -\infty$ sous M_2 .

Mais, plus que le comportement asymptotique, c'est celui à faible nombre de données qui nous intéresse ici. Son étude ne peut guère se faire qu'en simulation.

4.4 Comparaison expérimentale

Nous comparons les résultats obtenus par ce nouveau critère avec ceux fournis par des critères classiques sur un exemple simulé qui s'inspire directement de [5]. Le modèle « vrai » est caractérisé par $\boldsymbol{\mu} = \mathbf{X}_0\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_0 = [1, -2, 1/2, -1]^t$ ($p_0 = 4$). Deux situations sont étudiées. Dans la première, les éléments de la matrice \mathbf{X}_0 sont des variables i.i.d. et $\sim \mathcal{N}(0, \sigma_x^2)$, et le rapport signal-à-bruit est élevé (15 dB) (situation favorable). Les résultats présentés à la Fig. 1 montrent l'amélioration sensible apportée par le nouveau critère lorsque le nombre de données est faible.

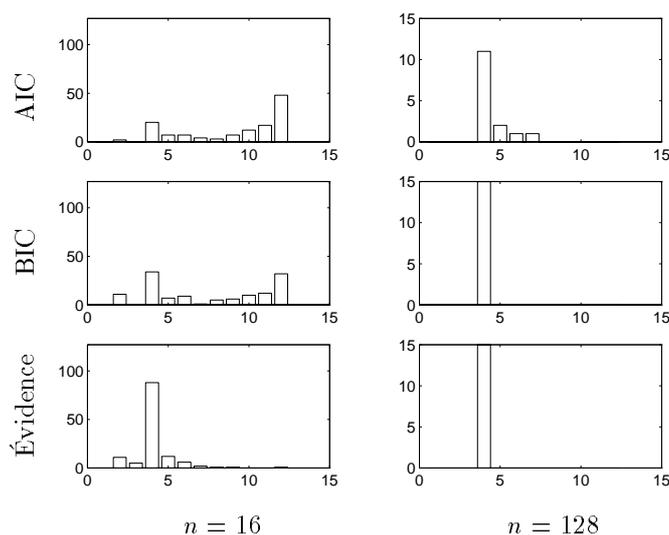


FIG. 1 - Histogrammes des ordres retenus sur un exemple de régression linéaire analogue à celui [5] ($p_0 = 4$, $p_{\max} = 12$). Les défauts du critère AIC sont évidents. Avec un nombre élevé de données ($n = 128$), les 2 autres méthodes fonctionnent bien, mais avec un faible nombre ($n = 16$), le critère BIC a une probabilité forte de nettement surévaluer l'ordre du modèle.

Dans la seconde situation, les variables gaussiennes qui constituent les éléments de la matrice \mathbf{X} sont fortement corrélées, et le rapport signal-à-bruit est faible (5 dB) (situation défavorable). Les résultats présentés à la Fig. 2 montrent que le nouveau critère apporte encore une amélioration lorsque le nombre de données est faible.

Références

- [1] H. Akaike, « A New Look at the Statistical Model Identification », *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716-723, 1974.
- [2] P. M. T. Broersen and H. E. Wensik, « Finite sample theory for autoregressive model estimation », *Actes du 13^e Colloque GRETSI*, Juan-les-Pins, 1991, pp. 297-300.
- [3] P. M. Djurić, « Asymptotic MAP Criteria for Model Selection », *IEEE Trans. on Signal Process.*, vol. 46, pp. 2726-2735, 1998.

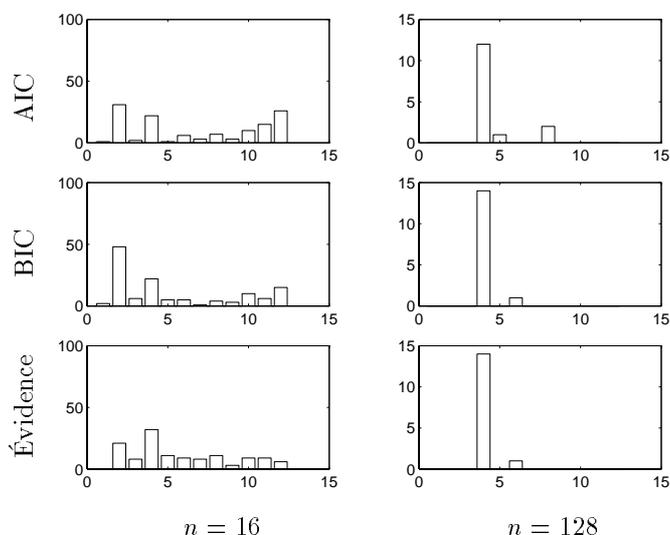


FIG. 2 - Exemple analogue à celui de la Fig. 1, mais cette fois la matrice \mathbf{X} est constituée de variables fortement corrélées et le rapport signal-à-bruit n'est que de 5 dB. Avec un faible nombre de données, l'amélioration apportée par le nouveau critère est encore sensible.

- [4] M. Evans and T. Swartz, « Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems », *Statistical Science*, vol. 10, pp. 254-272, 1995.
- [5] C. M. Hurvich and C.-L. Tsai, « Regression and time series model selection in small samples », *Biometrika*, vol. 76, pp. 297-307, 1989.
- [6] H. Jeffreys, *Theory of Probability*, Oxford: Clarendon Press, 1939, 5th augm. ed. 1988.
- [7] R. E. Kass and A. E. Raftery, « Bayes Factors », *J. Am. Statist. Assoc.*, vol. 90, pp. 773-795, 1995.
- [8] H. Linhart and W. Zucchini, *Model Selection*, New York: Wiley, 1986.
- [9] A. Nallanathan and W. J. Fitzgerald, « Bayesian model selection applied to spatial signal processing », *IEE Proc. Vis. Image Signal Process.*, vol. 141, pp. 76-80, 1994.
- [10] D. K. Pauler, « The Schwarz criterion and related methods for normal linear models », *Biometrika*, vol. 85, pp. 13-27, 1998.
- [11] J. Rissanen, « Modeling By Shortest Data Description », *Automatica*, vol. 14, pp. 465-471, 1978.
- [12] G. Schwarz, « Estimating the dimension of a model », *The Annals of Statistics*, vol. 6, pp. 461-464, 1978.
- [13] Y. Yang and A. R. Barron, « An Asymptotic Property of Model Selection Criteria », *IEEE Trans. Inform. Theory*, vol. 44, pp. 95-116, 1998.
- [14] A. Zellner and A. Siow, « Posterior odds ratio for selected regression hypothesis » (*with discussion*), in *Bayesian Statistics*, J.M. Bernardo *et al.*, eds., pp. 585-648, Valencia, Spain: University Press, 1980.