Intervalles de confiance pour une fonction implicite des paramètres d'un modèle : application au calcul de l'altitude optimale de présence d'espèces végétales dans une chaine montagneuse.

Vincent Couallier[1] & Audrey Eyermann[1] & Annabel J. Porté[2,3] & Morgane Urli[2,3]

(1) Institut Mathématique de Bordeaux UMR CNRS 5251, Université Victor Segalen
(2) INRA, UMR 1202 BIOGECO, F-33610 Cestas, France
(3) Université de Bordeaux, UMR 1202 BIOGECO, F-33400 Talence, France

Résumé : En couplant le théorème des fonctions implicites et la "méthode delta", on propose une méthode d'estimation de la variance d'un estimateur d'un paramètre défini implicitement comme le zéro d'une fonction $g(.,\theta)$ où θ est estimé par maximum de vraisemblance. L'application concerne la détermination de l'altitude optimale de présence d'espèces végétales le long d'un gradient d'altitude dans une chaine montagneuse.

Summary : By using a theorem of implicit functions and the Delta Méthod, we propose a method to estimate the variance of an estimator of an implicitly defined parameter such that the solution of $g(.,\theta) = 0$ where θ is estimated by the maximum likelihood method. We apply this to determine the altitude of the optimum of species presence on an altitudinal gradient in mountains.

Mots clés : fonctions implicites, intervalles de confiance, maximum de vraisemblance, delta méthode, optimum d'altitude.

1 Méthodes de calculs d'intervalles de confiance pour des fonctions implicites de paramètres d'un modèle statistique

On considère un modèle statistique paramétrique de paramètre $\theta \in \mathbb{R}^p$ et on se place dans le cadre d'une estimation par maximum de vraisemblance. Dans cette note, on propose une méthode de calcul d'intervalle de confiance pour une quantité $\phi \in \mathbb{R}$ qui est définie implicitement à partir de θ de la façon suivante : soit g une fonction continue à dérivée continue qui permet de définir implicitement $\phi \in \mathbb{R}^q$ par rapport à θ : $g(\phi, \theta) = 0$, en supposant que ϕ est défini de façon unique.

L'estimateur du maximum de vraisemblance du paramètre θ possède de nombreuses propriétés d'optimalité sous des hypothèses standarts. A partir de la normalité asymptotique de $\hat{\theta}_n$, l'objectif est d'obtenir la normalité (asymptotique) de $\hat{\phi}_n$ qui est solution

de $g(\hat{\phi}_n, \hat{\theta}_n) = 0$. De ce résultat, on peut déduire une construction d'intervalle de confiance. Cette méthode est comparée à un calcul d'intervalle de confiance par bootstrap non paramétrique.

1.1 Théorème des fonctions implicites et Delta-méthode

La "méthode delta" (voir Billingsley 1986, Serfling, 1980) permet de déduire d'un résultat de normalité (asymptotique le plus souvent) d'un estimateur $\hat{\theta}_n \in \mathbb{R}^p$, une approximation en loi d'une statistique de type $\hat{\phi}_n = f(\hat{\theta}_n)$, où $\hat{\phi}_n \in \mathbb{R}^q$, f est une fonction continuement différentiable de $\mathbb{R}^p$ dans $\mathbb{R}^q$. Si on connaît le gradient de f, évalué au point $\hat{\theta}_n$, et si on a un résultat du type

$$\Sigma_n^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}(0, Id_p)$$

où Σ_n est la matrice de variance asymptotique de $\hat{\theta}_n$, alors

$$(\nabla f' Var \nabla f)^{-1/2}_{|\theta=\hat{\theta}_n}(\hat{\phi}_n - \phi) \rightarrow_d \mathcal{N}(0, Id_q).$$

Couramment employée pour calculer la variance d'une transformation non linéaire des paramètres à partir de la normalité asymptotique d'un estimateur du maximum de vraisemblance, cette relation nécessite évidemment de connaître ou d'estimer la variance de $\hat{\theta}_n$ mais aussi d'expliciter le gradient de f. Or il existe des modèles où la quantité d'intérêt ϕ est définie implicitement à partir du paramètre θ, par exemple comme la racine d'une équation. L'objectif est donc de proposer, par une utilisation du théorème des fonctions implicites, une représentation du résultat ci-dessus dans le cas où ϕ est défini par $g(\phi, \theta) = 0$ où g est une fonction de $\mathbb{R}^{p+q}$ dans $\mathbb{R}^q$.

Le théorème des fonctions implicites (Taylor et Mann, 1983), affirme, sous certaines conditions, l'existence de fonctions continuement dérivables à valeurs réelles $f_1, \ldots, f_q$ dont on connait les dérivées partielles et qui sont telles que $g(\phi, \theta) = 0 \iff \phi = (f_1(\theta), \ldots, f_q(\theta))$. De plus, le gradient de $f = (f_1, \ldots, f_q)$ est connu :

$$\nabla f = \left[\frac{\delta f_i}{\delta \theta_j}\right] = -J^{-1}H$$

avec J la matrice carrée $q \times q$ des $(\frac{\delta g_i}{\delta \phi_j})$ et H la matrice $p \times q$ des $(\frac{\delta g_i}{\delta \phi_j})$.

Ce résultat, déjà utilisé dans Benichou et Gail (1989), permet d'obtenir une estimation de la variance d'un estimateur d'un paramètre défini implicitement à partir des paramètres du modèle, et donc un intervalle de confiance approché de type Wald puisque la variance asymptotique de $\hat{\phi}_n$ peut être calculée et vaut $J^{-1}H\Sigma_n H' J^{-1'}$ si Σ_n est la matrice de variance asymptotique de $\hat{\theta}_n$ (les dérivées étant évaluées au point $\hat{\theta}_n$).

Si le paramètre d'intérêt est la racine $\phi \in \mathbb{R}$ d'une fonction réelle $g(., \theta)$, racine supposée unique, le résultat prend la forme plus simple

$$\hat{S}^{-1/2}(\hat{\phi} - \phi) \rightarrow_d \mathcal{N}(0, 1)$$

où

$$\hat{S} = \frac{1}{\left(\frac{\delta g}{\delta \phi}(\hat{\phi}_n, \hat{\theta}_n)\right)^2} \left[\frac{\delta g}{\delta \theta_1}, \ldots, \frac{\delta g}{\delta \theta_p}\right]_{|(\phi,\theta)=(\hat{\phi}_n,\hat{\theta}_n)} \Sigma_n \begin{bmatrix} \frac{\delta g}{\delta \theta_1} \\ . \\ . \\ . \\ \frac{\delta g}{\delta \theta_p} \end{bmatrix}_{|(\phi,\theta)=(\hat{\phi}_n,\hat{\theta}_n).} \quad (1)$$

Un intervalle de confiance approché au niveau γ peut donc être obtenu par

$$IC_\gamma(\phi) = \left[\hat{\phi}_n - z_{\frac{1+\gamma}{2}}\sqrt{S}; \hat{\phi}_n + z_{\frac{1+\gamma}{2}}\sqrt{S}\right]$$

.

1.2 Intervalles de confiance par bootstrap

Une méthode alternative repose sur le ré-échantillonnage par bootstrap, soit des données de l'échantillon, soit du paramètre estimé $\hat{\theta}$ connaissant une estimation de sa matrice de variance. Que le paramètre d'intérêt ϕ soit défini implicitement ne pose pas de difficulté de calcul des intervalles de confiance par bootstrap (voir Davison et Hinkley 1997, pour plus de détails).

2 Application au calcul de l'optimum altitudinal de la présence d'espèces végétales

Les changements climatiques de ces dernières années préoccupent les écologistes de par leur rapidité et leur amplitude. En effet ils auraient des conséquences plus ou moins importantes sur les niches écologiques des espèces végétales, notamment dans les domaines montagneux. De nombreuses études ont montré que l'analyse du gradient altitudinal d'une espèce permettrait de mettre en évidence l'influence du facteur climatique sur les écosystèmes au cours du temps notamment en comparant la valeur de l'altitude où l'espèce est la plus abondante pour deux périodes données. C'est dans le but d'évaluer l'adaptation des espèces végétales aux modifications du facteur climatique que le laboratoire Biogeco analyse des données de présence de cinq espèces de feuillus au sein de deux gradients altitudinaux situés dans deux chaines de montagnes en Espagne (les Pyrénées et le Système Ibérique) et pour deux inventaires forestiers séparés d'une période de 10 ans. Ainsi pour chaque espèce étudiée, l'altitude des plots à laquelle l'espèce était présente ou absente a été notée par inventaire et chaine de montagne. L'objectif est d'analyser la répartition d'une espèce le long d'un gradient de température, mais également d'étudier l'évolution de la répartition altitudinale de l'espèce au cours du temps suivant le déplacement de sa niche écologique engendré par exemple par une augmentation de température dans le cadre du réchauffement climatique.

Les données, utilisées pour l'étude, correspondent à celles répertoriées lors de deux inventaires forestiers espagnols, qui ont eu lieu respectivement entre 1986 et 1996, et, 1997 et 2007 (notés SFI1990 et SFI2000). Chaque SFI correspond à un échantillonnage d'arbres effectué selon une grille systématique de placettes permanentes à travers l'Espagne au sein desquelles a été relevée la présence ou non des différentes espèces d'arbres. L'ensemble de la surface forestière est ainsi échantillonnée sur une grille carrée de 1 km de côté. Chaque placette est localisée par ses coordonnées géographiques UMT. Au total, 73772 et 67542 placettes sont suivies lors du SFI1990 et SFI2000 respectivement. Pour chaque inventaire, on a sélectionné deux zones d'étude où ont été effectués les relevés : le système ibérique et les Pyrénées. Pour chaque chaîne montagneuse a été notée l'altitude exacte de chaque placette où ont été observée la présence ou non de plusieurs espèces d'arbres.

2.1 le modèle asymétrique HOF V

Pour une année et une zone d'étude d'un inventaire donnés, on dispose d'un échantillon $(X_i, Y_i)_{i=1...n}$ où i est l'indice de la placette, X_i son altitude et Y_i l'indicateur de présence de l'espèce d'arbre à analyser. Différentes modélisations existent dans la littérature, les deux principales reposant sur un modèle de régression logistique et un modèle de présence-absence plus général, tous les deux utilisant l'altitude seule comme variable explicative de la présence d'une espèce d'arbre.

Pour obtenir une courbe de réponse unimodale, il est courant (Oksanen et al, 2001) de considérer un modèle classique de régression logistique

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(a+bx+cx^2)}}, \tag{2}$$

les coefficients a, b et c faisant de la courbe de probabilité de présence une courbe concave présentant un optimum en $alt_{opt} = -\frac{b}{2c}$. La répartition de présence autour de l'optimum d'altitude présentant une symétrie qui peut être difficilement compatible avec les données, Huisman et al (1993) ont proposé un modèle concurrent qui présente l'avantage de supporter des courbes de répartition symétrique et assymétrique. Le modèle $HOF\ V$ est

$$P(Y = 1|X = x) = \frac{1}{1 + e^{a+bx}} \frac{1}{1 + e^{c+dx}} \tag{3}$$

où les coefficients a, b, c et d peuvent être contraints à prendre certaines valeurs fixées ce qui définit les sous modèles HOF IV (b=-d) et HOF III (d=0).

Les modèles HOF V et HOF IV (comme la régression logistique (2)) sont adaptés à la modélisation de la courbe de probabilité de présence des espèces végétales selon un gradient d'altitude car ils peuvent admettre un optimum en une valeur d'altitude noté alt_{opt}. Ceci-dit, seul le modèle HOF V présente une courbe de réponse asymétrique autour de son optimum mais, celui-ci n'étant pas explicite, le problème est celui du calcul de la variance de l'estimateur $\widehat{alt_{opt}}$ pour déterminer un intervalle de confiance.

La méthode d'estimation repose sur la maximisation de la vraisemblance, on dispose donc d'une estimation de la matrice de variance de l'estimateur $\hat{\theta}$.

2.2 Détermination d'un intervalle de confiance pour alt_{opt}

On considère le modèle HOF V, pour lequel la valeur de l'altitude optimale (notée ϕ dans la suite) est l'unique solution de

$$\frac{d}{dx}P(Y=1|X=x) \propto g(x,\theta) = b\ e^{a+bx} + d\ e^{c+dx} + (b+d)\ e^{(a+c)+(b+d)x} = 0$$

Pour appliquer les résultats de la section précédente, on calcule :

$$\begin{aligned}
\frac{\delta g}{\delta a}|_{(\hat{\phi},\hat{\theta})} &= -\hat{d}e^{\hat{c}+\hat{d}\hat{\phi}} \\
\frac{\delta g}{\delta b}|_{(\hat{\phi},\hat{\theta})} &= e^{\hat{a}+\hat{b}\hat{\phi}} + e^{\hat{a}+\hat{c}+(\hat{b}+\hat{d})\hat{\phi}} - \hat{d}\hat{\phi}e^{\hat{c}+\hat{d}\hat{\phi}} \\
\frac{\delta g}{\delta c}|_{(\hat{\phi},\hat{\theta})} &= -\hat{b}e^{\hat{a}+\hat{b}\hat{\phi}} \\
\frac{\delta g}{\delta d}|_{(\hat{\phi},\hat{\theta})} &= e^{\hat{c}+\hat{d}\hat{\phi}} + e^{\hat{a}+\hat{c}+(\hat{b}+\hat{d})\hat{\phi}} - \hat{b}\hat{\phi}e^{\hat{a}+\hat{b}\hat{\phi}}
\end{aligned}$$

et

$$\frac{\delta g}{\delta \phi}|_{(\hat{\phi},\hat{\theta})} = -\hat{b}\hat{d}\left[e^{\hat{a}+\hat{b}\hat{\phi}} + e^{\hat{c}+\hat{d}\hat{\phi}}\right]. \tag{4}$$

Ainsi, connaissant une estimation $\hat{\Sigma}$ de la matrice de variance de $\hat{\theta} = (\hat{a},\hat{b},\hat{c},\hat{d})$, l'application de (1) permet d'obtenir une estimation de la variance de $\widehat{alt_{opt}}$ et donc le calcul d'un intervalle de confiance approchée en utilisant la normalité approchée issue de la "méthode delta".

Bibliographie

[1] Benichou, J. et Gail, M. (1989), A Delta Method for Implicitly Defined Random Variables, *The Amer. Statis.*, Vol. 43, No. 1 , pp. 41-44.

[2] Billingsley, P. (1986) *Probability and measure*, Wiley.

[3] Davison, A.C. et Hinkley, D.V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press.

[4] Huisman, J., Olff, H., Fresco, L.F.M. (1993) A hierarchical set of models for species response analysis, *Journal of Vegetation Science*, Vol 4, pp 37 - 46.

[5] Lawesson, J.E., Oksanen, J. (2002). Niche characteristics of Danish woody species as derived from coenoclines,*Journal of Vegetation Science*, Vol 13, pp 279 - 290.

[6] Lenoir, J., Gégout, J.C., Marquet, P.A., de Ruffray, P., Brisse, H. (2008) A significant upward shift in plant species optimum elevation during the 20th centur, *Science*, Volume 320, p. 1768.

[7] Oksanen, J., Laara, E., Tolonen, K. et Warner, B. (2001), Confidence intervals for the optimum in the gaussian response function, *Ecology*, Vol 82, pp 1191-1197.
[8] Serfling, R.J. (1980) *Approximation Theorems of mathematical Statistics*, New York, John Wiley.
[9] Taylor, A.E. et Mann, W.R. (1983) *Advanced Calculus* (3rd ed.), New York, John Wiley.