

# REVUE DE STATISTIQUE APPLIQUÉE

ANESTIS ANTONIADIS

JACQUES BERRUYER

ALAIN FILHOL

**Estimation semi-paramétrique dans les familles  
doublement poissonniennes et application aux  
spectres de diffraction**

*Revue de statistique appliquée*, tome 47, n° 3 (1999), p. 57-80

[http://www.numdam.org/item?id=RSA\\_1999\\_\\_47\\_3\\_57\\_0](http://www.numdam.org/item?id=RSA_1999__47_3_57_0)

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

# **ESTIMATION SEMI-PARAMÉTRIQUE DANS LES FAMILLES DOUBLEMENT POISSONNIENNES ET APPLICATION AUX SPECTRES DE DIFFRACTION**

Anestis Antoniadis\*, Jacques Berruyer†, Alain Filhol‡

\* *Université Joseph Fourier, LMC-IMAG, BP 53, 38041 Grenoble cedex 09, France*

† *Université Jean Monnet, UFR des Sciences, 23, rue du Dr. Paul Michelon,  
42100 St-Etienne Cedex, France.*

‡ *Institut Laue Langevin, BP 156, 38042 Grenoble cedex 9, France*

## **RÉSUMÉ**

Cet article fait la synthèse et étend les méthodologies statistiques utiles à l'analyse et l'étude des diagrammes de diffraction de poudre en cristallographie. Ce type de signal est la réalisation d'un processus de Poisson dont l'intensité est la somme d'une fonction lisse (bruit de fond) et d'un ensemble de fonctions paramétriques (pics de diffraction) dont il faut estimer les paramètres. Ceci est réalisé en modélisant les observations par un modèle de régression poissonnien non linéaire. En nous inspirant de la théorie des lois doublement exponentielles, développée par Efron [10], on introduit un paramètre supplémentaire permettant de contrôler la variance indépendamment de la moyenne. Enfin, le bruit de fond étant lisse, il est modélisé de manière non paramétrique par une combinaison de B-splines. Afin de séparer au mieux les pics du fond et de réduire l'influence de l'estimation de ce dernier sur l'estimation des paramètres des pics, les coefficients de cette combinaison sont pénalisés de manière optimale par des techniques de validation croisée.

*Mots-clés : Familles doublement poissonniennes, Modèles généralisés, Surdispersion, Sous-dispersion, B-splines, Validation croisée, Spectres de diffraction.*

## **ABSTRACT**

This paper synthetically presents and extends the statistical methods available for analysing and studying powder diffraction spectra in Crystallography. Signals of this form can be considered as a sample from a Poisson process the intensity of which can be modelled as a sum of a smooth component (the background signal) and a collection of parametric functions (the diffraction peaks) with unknown parameters. To estimate the background and the peak parameters, a nonlinear Poisson regression is carried out. When analysing such data, a habitual source of concern is the presence of overdispersion. Double exponential families, developed by Efron [10], allow the introduction of an extra parameter that controls variance independently

of the mean. The background signal being smooth, we propose to estimate it with a B-spline smoother. To better separate the peaks from the background and in order to minimize the effects of the smoothing procedure on the peaks's parameter estimates, a difference penalty on adjacent B-spline coefficients is incorporated into a penalised version of Fisher scoring algorithm and the penalty is optimally regulated by cross-validation.

**Keywords :** *Double Poisson families, Generalized models, overdispersion, underdispersion, B-splines, cross-validation, Diffraction spectra.*

## 1. Introduction et motivation

Cet article fait la synthèse et étend les méthodologies statistiques utiles à l'analyse et l'étude des diagrammes de diffraction de poudre en cristallographie. Ce type de diagrammes, classiquement enregistrés à l'aide de rayons X ou de neutrons, est très couramment utilisé pour la caractérisation d'échantillons purs ou composites, pour l'étude de transition de phases ou pour la détermination de la structure de composés à petites molécules [3].

Un diagramme de diffraction de poudre est un ensemble de  $n$  points mesurés,  $(t_i, N_i)$ ,  $i = 1, \dots, n$ , où  $t_i$  désigne un angle de diffraction (ou angle de Bragg) et  $N_i$  est le résultat d'un comptage de particules élémentaires (photons ou neutrons) diffractées à cet angle. L'ensemble des comptages observés est supposé être la réalisation d'un modèle de régression poissonnienne avec, pour variable explicative, l'angle de diffraction. L'aspect général d'un diagramme de diffraction de poudre est décrit par une ligne de base de forme quelconque mais assez lisse (bruit de fond) auquel sont superposés des pics de diffraction d'intensités et de formes variables. Le bruit de fond est en général composé d'une somme de signaux dûs à des phénomènes physiques autres que la diffraction (rayonnement d'ambiance, diffusion diffuse, etc...). Ce bruit de fond généralement assez homogène peut néanmoins présenter de faibles oscillations (désordre partiel, fluctuations au voisinage d'une transition,...). Par contre, les pics sont très localisés, et peuvent être le plus souvent représentés par des fonctions paramétriques. Ils peuvent être bien séparés ou très nombreux avec un fort taux de recouvrement, de formes similaires ou très différentes (mélange de composés, sous réseaux cristallographiques de natures différentes, etc...).

Dans la majorité des cas l'information utile au cristallographe est portée par les pics. Leurs positions et intensités sont directement liées à la structure atomique du matériau tandis que leur largeur dépend, pour l'essentiel, de caractéristiques instrumentales et de caractéristiques macroscopiques de l'échantillon (forme, mosaïque,...). En l'absence d'hypothèse, d'une part, sur la structure du matériau échantillon responsable de la diffraction (pics) et, d'autre part, sur l'influence de paramètres instrumentaux (bruit de fond), on fait appel à un modèle purement descriptif ne nécessitant de connaissance *a priori*, ni sur l'échantillon diffracté, ni sur l'instrument de mesure. Dans ce cas le signal est supposé être la somme d'une fonction lisse et d'un ensemble de fonctions paramétriques (pics de diffraction) dont il faut estimer les paramètres le plus efficacement possible.

Pour l'analyse inférentielle de données de diffraction il est naturel de considérer un modèle de régression poissonnien, étendant dans un cadre non linéaire les modèles linéaires généralisés, introduits par Nelder et Wedderburn [14] pour l'analyse de données issues d'une famille exponentielle. Pour ce type de modèles, linéaires ou non,

la variance est une fonction de la moyenne, ce qui en pratique restreint considérablement le champ d'application. En effet, le processus d'acquisition et le pré-traitement des mesures (chaîne de comptage, calibration de multidétecteurs) dénature l'aléa poissonnien du flux de particules. En nous inspirant de la théorie des lois doublement exponentielles, développée par Efron [10], nous nous proposons d'introduire un paramètre supplémentaire permettant de contrôler la variance indépendamment de la moyenne. Ce type de familles exponentielles courbées que nous appellerons *familles doublement poissonniennes* permettra une analyse statistiquement correcte des données de diffraction, tenant compte de phénomènes de sur ou sous-dispersion par rapport à la loi de Poisson.

Pour tenir compte d'un bruit de fond lisse, mais de forme complexe, ce dernier sera modélisé par une combinaison de B-splines. Afin de séparer au mieux le bruit de fond des pics et de réduire l'influence de son estimation sur celle des paramètres des pics, les coefficients de cette combinaison seront pénalisés de manière optimale.

Nous rappelons brièvement dans le paragraphe 2, les définitions et les résultats des modèles de régression poissonnienne non linéaires, dérivés de résultats plus classiques sur les modèles linéaires généralisés. Dans le paragraphe 3 nous présentons les familles doublement poissonniennes et nous précisons les propriétés des estimateurs du maximum de vraisemblance des paramètres ainsi que l'algorithme de résolution numérique associé, lorsque le bruit de fond est supposé être de forme paramétrique donnée. Le paragraphe 4 est consacré à une modélisation semi-paramétrique du signal à estimer à l'aide d'une paramétrisation de type B-spline du bruit de fond. Les performances des estimateurs sont étudiées dans le paragraphe 5 à l'aide de simulations et une analyse de données réelles conclut cet article.

## 2. Modèles de régression non linéaire poissonnienne

Dans ce paragraphe nous rappelons brièvement les notions importantes sur les modèles généralisés non linéaires qui nous seront utiles par la suite. Pour plus de détails le lecteur pourra se reporter au chapitre V de [3] ou à l'article [4].

Pour ces modèles de régression, la réponse  $Y(t)$  est issue d'une loi appartenant à une structure exponentielle dont le paramètre naturel  $\alpha$  est une fonction paramétrique du régresseur  $t$ , notée  $f(\theta, t)$ . Plus précisément, la loi de probabilité régissant les observations est un élément de la famille exponentielle de lois de probabilité sur  $\mathbb{R}$  définie par

$$\mathcal{G} = \{G_\alpha, \alpha \in M\}$$

avec  $M$  un ouvert d'intérieur non vide de  $\mathbb{R}$ ,

$$\frac{dG_\alpha}{dG_0}(y) = g_{\alpha,1}(y) = e^{\alpha y - \psi(\alpha)}$$

où  $\alpha = f(\theta, t)$ ,  $\theta \in \mathbb{R}^p$  et  $G_0$  est une mesure dominante de  $\mathcal{G}$ ,  $y$  est la statistique canonique et  $\psi$  est la fonction cumulée (voir par exemple Barndorff-Nielsen [5]). Sous des conditions classiques de régularité, il est souvent commode de paramétrer une famille exponentielle par sa moyenne plutôt que par son paramètre naturel. Ainsi,

par exemple, dans le cas des données régies par une loi poissonnienne paramétrée par sa moyenne on a :

$$\frac{dG_\mu}{dG_0}(y) = g_{\mu,1}(y) = e^{\log(\mu)y - \mu}, \quad (1)$$

avec  $\mu = \exp(\alpha) = \psi(\alpha)$ . Lorsque  $f(\theta, t)$  est une fonction linéaire en  $\theta$ , on retrouve en particulier les modèles linéaires généralisés avec pour fonction de lien, la fonction de lien canonique. Dans le cas contraire, le famille exponentielle correspondante est une famille courbe au sens de Efron (voir par exemple [3]).

Plus généralement, on notera par la suite  $g_{\alpha,n}(y)$  et  $g_{\mu,n}(y)$  les fonctions définies par

$$g_{\alpha,n}(y) = e^{n[\alpha y - \psi(\alpha)]} \quad \text{et} \quad g_{\mu,n}(y) = e^{n[\log(\mu)y - \mu]}.$$

### *Estimation du maximum de vraisemblance dans les familles poissonniennes*

Soit  $Y_1, \dots, Y_n$  une suite de variables aléatoires indépendantes de lois ayant pour densités respectives  $g_{\eta_i(\beta),1}$  où  $\eta(\beta)$  est une fonction donnée de  $\mathbb{R}^p$  à valeurs dans  $\mathbb{R}^n$ , supposée satisfaire aux hypothèses habituelles de la moyenne d'une régression non linéaire. Les composantes  $\eta_i(\beta)$  ( $i = 1, \dots, n$ ) du vecteur  $\eta(\beta)$ , qu'on notera aussi  $\eta_i$ , sont les paramètres de la loi de Poisson associée à  $Y_i$ .

La fonction log de vraisemblance  $\mathcal{L}(\beta)$  associée à une réalisation  $y_1, \dots, y_n$  de la suite  $Y_1, \dots, Y_n$  est donc égale à :

$$\mathcal{L}(\beta) = \sum_{i=1}^n [y_i \log(\eta_i) - \eta_i].$$

La recherche des estimateurs du maximum des vraisemblance passe par le calcul du vecteur des scores  $U(\beta) = \frac{\partial \mathcal{L}}{\partial \beta}$ . Un calcul simple donne :

$$U(\beta) = {}^t \partial \eta / \partial \beta \cdot \partial \mathcal{L} / \partial \eta = {}^t X(\beta) \partial \mathcal{L} / \partial \eta$$

où

$$X(\beta) = \frac{\partial \eta}{\partial \beta}.$$

Notons également que :

$$\frac{{}^t \partial \eta}{\partial \beta} \frac{\partial \mathcal{L}}{\partial \eta} = {}^t X(\beta) \left[ \frac{y_i}{\eta_i} - 1 \right],$$

où  $\left[ \frac{y_i}{\eta_i} - 1 \right]$  est le vecteur  $n$ -dimensionnel de composantes  $\left( \frac{y_i}{\eta_i} - 1 \right)$ .

Les estimateurs du maximum de vraisemblance annulent le vecteur des scores. Le système à résoudre étant non linéaire, on applique, pour la recherche des

estimateurs, la méthode des scores de Rao [15]. Il faut pour cela calculer la matrice d'information selon Fisher associée à la fonction de vraisemblance. Cette dernière n'est autre que la matrice de variance-covariance du vecteur des scores. Pour cela il faut d'abord calculer l'information associée au paramètre  $\eta$ . Il est facile de voir que cette dernière est une matrice diagonale d'ordre  $n$  :

$$\mathcal{I}_n = (\text{var}(\partial\mathcal{L}/\partial\eta)) = \text{Diag} [\text{var}(Y_i)/\eta_i^2] = \text{Diag} [1/\eta_i],$$

où  $\text{Diag}[\xi_i]$  désigne une matrice diagonale avec  $\xi_i$  pour éléments diagonaux.

Nous pouvons maintenant définir l'algorithme itératif de type descente pour obtenir des estimations du maximum de vraisemblance des paramètres inconnus. L'étape  $r$  de l'algorithme des scores sera :

$$\beta^{r+1} = \beta^r + \alpha_r M^{-1} X(\beta^r) \partial\mathcal{L}_r / \partial\eta$$

où

$$M = \mathbb{E} [-\partial^2\mathcal{L} / \partial\beta^t \partial\beta],$$

est la matrice d'information théorique associée au modèle.

On trouve pour  $M$  l'expression suivante :

$$M = {}^t X(\beta) \text{Diag} [1/\eta_i] X(\beta).$$

### 3. Familles doublement poissonniennes

Comme nous l'avons déjà souligné, les procédures d'acquisition des données de diffraction dénaturent souvent le caractère poissonnien des observations en altérant la dispersion des données. En effet, l'utilisation des multidétecteurs statiques (chaque  $y_i$  est mesuré par une cellule différente du détecteur) à la place d'un détecteur unique mobile mesurant les  $y_i$  de manière séquentielle introduit une hétérogénéité des réponses des différentes cellules du détecteur. Pour en tenir compte un coefficient multiplicatif  $\alpha_i$  est appliqué aux mesures brutes  $y_i$  altérant le caractère poissonnien des observations. La théorie des lois doublement exponentielles nous permet d'introduire un paramètre supplémentaire pour contrôler la variance indépendamment de la moyenne. Ce type de familles exponentielles courbées, que nous appellerons familles doublement poissonniennes, nous permettra une analyse statistiquement correcte des données de diffraction, tenant compte de phénomènes de sur ou sous-dispersion par rapport à la loi de Poisson.

Les familles des lois doublement exponentielles furent introduites pour la première fois par Diaconis et Efron [9] pour l'étude des tests d'indépendance du  $\chi^2$  dans les tables de contingence. Elles sont définies, comme pour le cas des familles exponentielles classiques, par la donnée de leur densité par rapport à une mesure dominante  $G_0$ , c'est-à-dire, par :

$$\frac{dG_{\mu,\gamma}}{dG_0}(y) = \tilde{f}_{\mu,\gamma}(y) = \phi(\mu, \gamma) g_{\mu,1}(y) \left( \frac{g_{\mu,1}(y)}{g_{y,1}(y)} \right)^{\gamma-1}, \quad \gamma \in ]0, 1],$$

où  $\phi(\mu, \gamma)$  est une constante de normalisation et  $g_{\mu,1}$  est la fonction définie au début du paragraphe 2. On peut s'apercevoir que pour  $\gamma$  fixé, et à condition de changer de mesure dominante, la famille des densités précédentes, paramétrées par  $\mu$ , est une famille de lois exponentielles au sens classique. En effet on a :

$$\tilde{f}_{\mu,\gamma}(y)dG_0(y) = K_{\mu,\gamma}e^{\gamma\alpha(\mu)y - \psi(\alpha(\mu))\gamma}dG'_\gamma(y),$$

avec

$$K_{\mu,\gamma} = \phi(\mu, \gamma)$$

et

$$dG'_\gamma(y) = [\exp(\alpha(y)y - \psi(\alpha(y)))]^{1-\gamma} dG_0(y).$$

Par contre lorsque l'on traite  $\gamma$  comme paramètre variable, cette famille n'est plus une famille exponentielle.

Un calcul simple montre que la densité par rapport à la mesure produit  $dG'_\gamma^{\otimes n}$  d'un échantillon de taille  $n$  issu d'une loi doublement exponentielle peut s'écrire sous la forme :

$$\phi_n(\mu, \gamma)g_{\mu,n}(y)^\gamma g_{y,n}(y)^{1-\gamma},$$

où  $y$  désigne la moyenne empirique de l'échantillon et  $\phi_n(\mu, \gamma)$  est la constante normalisante. Remarquons que pour  $\gamma = 1$  on trouve comme densité  $\phi_n(\mu, 1)g_{\mu,n}(y)$  et donc  $\phi_n(\mu, 1) = 1$ .

Une méthode d'estimation de la constante de normalisation  $\phi_n(\mu, \gamma)$  est développée dans l'article de Diaconis et Efron [9]. Elle est fondée sur la représentation de Hoffding d'une famille exponentielle et sur des approximations de type tilda classiques pour développement d'Edgeworth dans le cadre des familles exponentielles (voir Barndorff-Nielsen et Cox [6], p. 181). Le lecteur intéressé pourra se reporter à Efron [10]. Citons simplement le résultat :

$$\phi_n(\mu, \gamma) = \sqrt{\gamma}[1 + \mathcal{O}((n\gamma)^{-1})].$$

Dans le cas particulier où la famille exponentielle de départ est une loi de Poisson de paramètre  $\mu$ , on a :

$$g_{\mu,1}(y) = e^{y \log(\mu) - \mu} dG_0(y), \quad \alpha(\mu) = \log(\mu), \quad \psi(\alpha(\mu)) = \mu,$$

où  $G_0$  est la mesure dénombrement avec une masse  $1/y!$  en  $y$  entier. Puisque,  $t = ny = n\bar{y}_n$  suit une loi de Poisson de paramètre  $n\mu$ , on obtient :

$$g_{\mu,n}(y) = ne^{ny \log(n\mu) - n\mu} dG_0(ny) = e^{n[y \log(\mu) - \mu]} dH_n(y),$$

avec

$$dH_n(y) = n^{ny+1} dG_0(ny).$$

On peut donc définir la famille de lois doublement exponentielles associée, que nous appellerons famille de lois doublement poissonniennes. Pour cela, comme dans le cas général on pose :

$$\begin{aligned} \tilde{f}_{\mu,\gamma,n}(y) &= \phi_n(\mu, \gamma) g_{\mu,n}^\gamma(y) g_{y,n}^{1-\gamma}(y) dH_n(y) \\ &= c(\mu, \gamma, n) \sqrt{\gamma} g_{\mu,1}^{n\gamma}(y) g_{y,1}^{n-n\gamma}(y) dH_n(y), \end{aligned}$$

avec  $c(\mu, \gamma, n) = \phi_n(\mu, \gamma) / \sqrt{\gamma}$ . Comme l'asymptotique  $n = 1$  et  $\mu \rightarrow \infty$  équivaut approximativement, dans le cas poissonnien, à fixer la valeur de  $\mu$  et faire tendre  $n$  vers l'infini, on notera plus généralement la constante de normalisation  $c(\mu, \gamma, n)$  par  $c(\mu, \gamma)$ . En utilisant l'approximation tilda développée dans le cadre général des familles doublement exponentielles on trouve que cette constante de normalisation  $c(\mu, \gamma)$  est telle que :

$$c(\mu, \gamma) = 1 + \mathcal{O}((\mu\sqrt{\gamma})^{-1}),$$

quand  $\mu$  tend vers l'infini. Une approximation  $c_{app}$  plus précise de la constante  $c(\mu, \gamma)$  dans le cas poissonnien est donnée par Efron dans [10] :

$$\frac{1}{c_{app}(\mu, \gamma)} = 1 + \frac{1-\gamma}{12\mu\gamma} \cdot \left(1 + \frac{1}{\mu\gamma}\right).$$

La qualité de cette approximation est discutée dans Aitkin [2] et Seeber [17].

Pour apprécier la qualité de cette approximation nous avons calculé la valeur de  $c(\mu, \gamma)$  pour diverses valeurs de  $\mu$  et de  $\gamma$ . Les résultats sont reportés dans le tableau ci-dessous.

TABLEAU 0  
Rapport  $c(\mu, \gamma)/c_{app}(\mu, \gamma)$   
des constantes de normalisation et de leur approximation  
pour la famille doublement poissonnienne.

		1	0.75	$\gamma$ 0.5	0.333
$\mu$	10	1	0.9998	0.9985	0.9961
	5	1	0.9984	0.997	1.0125
	2	1	1.0048	1.0539	1.1979
	1	1	1.0529	1.2567	1.7592
	0.5	1	1.2170	1.9435	3.8005

Les résultats précédents nous permettent, pour des valeurs de  $\mu$  assez grandes (plus grandes que 2) et pour  $n = 1$  de remplacer la densité doublement poissonnienne  $\tilde{f}_{\mu,\gamma,n}(y)$  par la «densité» approximative  $f_{\mu,\gamma}(y)$  définie par :

$$f_{\mu,\gamma}(y) = \sqrt{\gamma} e^{-\mu\gamma} e^{-y} \frac{y^y}{y!} \left(\frac{e\mu}{y}\right)^{\gamma y}.$$

On utilisera cette densité approximative pour les calculs.



*Paramètre de sur ou sous-dispersion*

Soit  $Y$  une variable aléatoire de loi doublement poissonnienne, dont nous approcherons la densité par la fonction  $f_{\mu,\theta}(y)$  donnée par :

$$f_{\mu,\theta}(y) = \sqrt{\theta} e^{-\mu\theta} e^{-y} \frac{y^y}{y!} \left( \frac{e\mu}{y} \right)^{\theta y}.$$

Effectuons, dans l'expression précédente l'approximation de Stirling

$$y! \sim \sqrt{2\pi} y^{y+1/2} e^{-y}.$$

On trouve

$$f_{\mu,\theta}(y) \sim \sqrt{\theta} e^{-\mu\theta} \frac{1}{\sqrt{2\pi}} y^{-\theta y - 1/2} (e\mu)^{\theta y}.$$

La formule de Stirling dans le cas continu

$$\sqrt{2\pi} (\theta y)^{\theta y + 1/2} e^{-\theta y} \sim \Gamma(\theta y + 1),$$

donne comme approximation de  $f_{\mu,\theta}(y)$  la densité

$$\theta \frac{1}{\Gamma(\theta y + 1)} e^{-\mu\theta} (\mu\theta)^{\theta y}.$$

D'après Abrahamovitz et Stegun [1], une variable aléatoire de densité

$$g_{\mu,\theta}(x) = \frac{1}{\Gamma(x + 1)} e^{-\mu\theta} (\mu\theta)^x,$$

admet pour espérance  $\theta\mu$  et pour variance  $\theta\mu$ . On en déduit donc la propriété suivante :

**Proposition.** *Si  $X$  est une variable aléatoire de densité  $g_{\mu,\theta}$  alors la loi de la variable aléatoire  $Y = X/\theta$  est, pour  $\mu$  tendant vers l'infini, approximativement une loi doublement poissonnienne et on a  $\mathbb{E}(Y) = \mu$  et  $\text{var}(Y) = \mu/\theta$ .*

La proposition précédente permet donc d'interpréter le paramètre  $\theta$  comme un paramètre de sous ou sur-dispersion par rapport à une loi de Poisson de paramètre  $\mu$ .

La densité  $f_{\mu,\theta}(y)$  de la loi de  $Y$  par rapport à la mesure dénombrement n'est pas la densité d'une loi exponentielle dispersée au sens de Jorgensen [13]. En effet il est montré par cet auteur qu'une famille de lois exponentielles dispersées ne peut, lorsqu'elle est discrète, être dominée par une mesure dominante commune, *i.e.* dont le support est indépendant des paramètres.

#### 4. Estimation du maximum de vraisemblance dans les familles doublement Poissonniennes

Soit  $Y_1, \dots, Y_n$  une suite de variables aléatoires indépendantes de lois ayant pour densités respectives  $f_{\eta_i(\beta), \lambda \rho_i}$  où  $\eta(\beta)$  est une fonction donnée de  $\mathbb{R}^p$  à valeurs dans  $\mathbb{R}^n$ , supposée satisfaire aux hypothèses de régularité habituelles sur la moyenne dans le cadre d'une régression non linéaire ([3]). Les paramètres  $\lambda > 0$  et  $\rho_i$  sont des paramètres de dispersion de nature différente :  $\lambda$  est un paramètre de dispersion général, alors que les  $\rho_i$  sont des constantes de dispersion connues, introduites lors du processus d'acquisition des mesures par des transformations d'échelle sur les données.

La fonction log de vraisemblance  $\mathcal{L}(\beta, \lambda)$  associée à une réalisation  $y_1, \dots, y_n$  de la suite  $Y_1, \dots, Y_n$  est donc égale, à une constante près ne dépendant que des  $y_i$  et des  $\rho_i$ , à :

$$\mathcal{L}(\beta, \lambda) = \frac{n}{2} \log(\lambda) + \lambda \sum_{i=1}^n [\rho_i y_i \log(\eta_i) - \rho_i \eta_i] - \lambda \sum_{i=1}^n [\rho_i y_i \log(y_i) - \rho_i y_i].$$

La recherche des estimateurs du maximum de vraisemblance passe par le calcul du vecteur des scores

$$U(\beta, \lambda) = {}^t \left[ \frac{\partial \mathcal{L}}{\partial \beta} \quad \frac{\partial \mathcal{L}}{\partial \lambda} \right].$$

Un calcul simple donne :

$$U(\beta, \lambda) = \left[ \begin{array}{c} {}^t \partial \eta / \partial \beta \quad \partial \mathcal{L} / \partial \eta \\ n / (2\lambda) - \sum_{i=1}^n (\rho_i y_i \log(y_i / \eta_i) - \rho_i (y_i - \eta_i)) \end{array} \right] = \left[ \begin{array}{c} {}^t X(\beta) \partial \mathcal{L} / \partial \eta \\ \partial \mathcal{L} / \partial \lambda \end{array} \right]$$

où

$$X(\beta) = \frac{\partial \eta}{\partial \beta}.$$

Notons également que

$$\frac{\partial \eta}{\partial \beta} \frac{\partial \mathcal{L}}{\partial \eta} = {}^t X(\beta) \text{Diag} \left[ \lambda \rho_i \left( \frac{y_i}{\eta_i} - 1 \right) \right].$$

Les estimateurs du maximum de vraisemblance annulent le vecteur des scores. Le système à résoudre étant non linéaire, nous nous proposons d'appliquer, pour la recherche des estimateurs, la méthode des scores de Rao. Il faut pour cela calculer la matrice d'information selon Fisher associée à la fonction de vraisemblance. Cette dernière n'est autre que la matrice de variance-covariance du vecteur des scores. On

calculera d'abord l'information partielle associée au paramètre  $\eta$ . Il est facile de voir que cette dernière est une matrice diagonale d'ordre  $n$  :

$$\mathcal{I}_n = \text{Diag} [\lambda^2 \rho_i^2 \text{var}(Y_i) / \eta_i^2] = \text{Diag} [\lambda \rho_i / \eta_i].$$

Nous pouvons maintenant définir l'algorithme itératif de type descente pour obtenir des estimations du maximum de vraisemblance des paramètres inconnus. L'étape  $r$  de l'algorithme des scores sera :

$$\begin{bmatrix} \beta^{r+1} \\ \lambda^{r+1} \end{bmatrix} = \begin{bmatrix} \beta^r \\ \lambda^r \end{bmatrix} + \alpha_r M^{-1} \begin{bmatrix} {}^t X(\beta^r) \partial \mathcal{L}_r / \partial \eta \\ \partial \mathcal{L}_r / \partial \lambda \end{bmatrix}$$

où

$$M = \mathbb{E} \begin{bmatrix} -\partial^2 \mathcal{L} / \partial \beta^t \partial \beta & -\partial^2 \mathcal{L} / \partial \beta \partial \lambda \\ -\partial^2 \mathcal{L} / \partial \lambda \partial \beta & -\partial^2 \mathcal{L} / \partial \lambda^2 \end{bmatrix}$$

et  $\mathcal{L}_r$  désigne la fonction Log de vraisemblance calculée en  $(\beta^r, \lambda^r)$ .

On trouve pour  $M$  l'expression suivante :

$$M = \begin{bmatrix} {}^t X(\beta) \text{Diag} [\lambda \rho_i / \eta_i] X(\beta) & 0 \\ 0 & n / (2\lambda^2) \end{bmatrix}$$

Dans le calcul précédent, nous avons pris pour  $M$  la matrice d'information théorique associée au modèle. Le fait que la matrice d'information est diagonale par bloc entraîne que les estimateurs du maximum de vraisemblance des paramètres  $\lambda$  et  $\beta$  sont asymptotiquement indépendants.

Comme dans le cas de la régression non linéaire sur des structures exponentielles, deux autres choix sont possibles pour  $M$ . Le premier consiste à prendre la matrice d'information observée, mais cette dernière n'étant pas une matrice diagonale par bloc cela entraînerait des calculs trop lourds pour que cette méthode ait un intérêt pratique. Le deuxième consiste à linéariser la matrice d'information observée pour obtenir une matrice  $M'$  bloc-diagonale constituée de deux blocs, l'un d'ordre  $n$  et l'autre d'ordre 1. Plus précisément :

$$M' = \begin{bmatrix} {}^t X(\beta) \text{Diag} [\lambda \rho_i y_i / \eta_i^2] X(\beta) & 0 \\ 0 & n / (2\lambda^2) \end{bmatrix}.$$

Cette dernière est en général semi-définie positive alors que la matrice d'information théorique est toujours définie positive; c'est la raison principale pour laquelle nous retenons pour la suite la matrice  $M$ , l'algorithme d'optimisation associé étant la méthode des scores de Rao [15].

Remarquons d'abord que le vecteur des  $n$  premières composantes du vecteur des scores  $U(\beta, \lambda)$  correspondant au gradient de la fonction de vraisemblance relativement à  $\beta$  est, au facteur multiplicatif  $\lambda$  près, égal au vecteur des scores

d'une structure doublement poissonnienne pour laquelle le paramètre de sur ou sous-dispersion  $\lambda$  vaut 1. Plus précisément, on a :

$$\frac{\partial \mathcal{L}}{\partial \beta} = \lambda {}^t X(\beta) \text{Diag} \left[ \frac{\rho_i y_i}{\eta_i} - \rho_i \right].$$

Remarquons ensuite que le bloc  $[M]_p$  d'ordre  $p$  de la matrice  $M$  est également proportionnel, au facteur multiplicatif  $\lambda$  près, à la matrice d'information du modèle sans facteur  $\lambda$ , *i.e.* :

$$[M]_p = {}^t X(\beta) \text{Diag} [\rho_i / \eta_i] X(\beta).$$

Comme cette matrice intervient par son inverse dans l'algorithme itératif de mise à jour de l'estimateur de  $\beta$ , on voit aisément que l'estimateur de  $\beta$  obtenu par la méthode des scores est indépendant de la valeur de  $\lambda$ . L'estimation du maximum de vraisemblance de  $\beta$  est donc identique à celle que l'on aurait obtenue en supposant que le facteur  $\lambda = 1$ . Si  $\hat{\beta}$  désigne cette estimation, obtenue à la convergence de la méthode de descente, la dernière composante du vecteur des scores conduit à l'estimation :

$$\hat{\lambda} = \frac{n}{\text{Dev}(y, \eta(\hat{\beta}))}$$

où

$$\text{Dev}(y, \eta) = 2 \sum_{i=1}^n \rho_i (y_i \log(y_i / \eta_i) - y_i + \eta_i)$$

est la déviance du modèle doublement poissonnien sans facteur de sur ou sous-dispersion.

La matrice  $M'$  présentant une structure similaire à celle de  $M$ , l'algorithme précédent s'adapte facilement, mais il faudrait garantir que  $M'$  reste définie positive lors des itérations, ce qui est automatique pour  $M$ . Enfin, la simplicité de l'algorithme justifie, encore une fois, le choix d'une matrice d'information théorique plutôt que d'une matrice d'information observée.

### Résultats asymptotiques et adéquation

Sous les hypothèses habituelles de régularité dans les modèles de régression non linéaire pour les familles exponentielles (voir [3]), les estimateurs du maximum de vraisemblance  $\hat{\beta}$  et  $\hat{\lambda}$  sont fortement consistants et stochastiquement indépendants. De plus  $\hat{\beta}$  suit asymptotiquement une loi gaussienne de moyenne  $\beta$  et de matrice de variance-covariance l'inverse du bloc supérieur d'ordre  $n$  de la matrice d'information théorique. Enfin, la variable aléatoire :

$$\frac{\text{Dev}(y, \eta(\hat{\beta}))}{\lambda}$$

suit asymptotiquement une loi du  $\chi^2$  à  $n - p$  degrés de liberté.

Au vu de ce dernier résultat, pour estimer par maximum de vraisemblance le paramètre  $1/\lambda$ , nous prendrons l'estimateur  $1/\hat{\lambda}$ . L'estimateur ainsi obtenu a, du moins asymptotiquement, une espérance égale à  $(n-p)/(n\hat{\lambda})$  et on préférera utiliser comme estimateur de  $1/\lambda$  la statistique :

$$1/\hat{\lambda} = \frac{\text{Dev}(y, \eta(\hat{\beta}))}{(n-p)}$$

asymptotiquement non biaisée et de variance asymptotique donnée par :

$$\text{var}(1/\hat{\lambda}) = 2/[(n-p)\lambda^2].$$

### 5. Modèle semi-paramétrique

Dans l'ajustement de spectres de diffraction de poudre, les données sont des réalisations  $y_i$  d'une suite de variables aléatoires indépendantes doublement poissonniennes dont le paramètre est de la forme :

$$\mu_i = B(t_i) + \sum_{k=1}^p c_k f_k(t_i, \theta_k)$$

où

- $B(t)$  est une fonction de l'angle de diffraction, supposée lisse et modélisant le bruit de fond;
- les fonctions  $f_k$  sont des fonctions paramétriques modélisant les  $p$  pics de diffraction. Le nombre de pics  $p$  est en général connu. Pour l'exemple que nous avons traité et pour les simulations illustrant notre méthode, les pics seront des densités gaussiennes, paramétrées soit par leur moyenne ( $\mu$ ) et leur écart-type ( $\sigma$ ), soit par leur moyenne et leur largeur à mi-hauteur ( $L = 2\sigma\sqrt{2\log(2)}$ ) comme le font habituellement les cristallographes.

Dans [4] le bruit de fond  $B(t)$  est modélisé par un polynôme de degré inférieur ou égal à 2. Lorsque cette approximation polynomiale du bruit de fond est raisonnable, un ajustement correct du polynôme permet d'obtenir des estimations des paramètres des pics pratiquement non biaisées. Il arrive pourtant qu'une étude exploratoire des données de diffraction montre que les modèles de ce type sont inadaptés car le bruit de fond, bien que lisse, ne peut être modélisé globalement par un polynôme de faible degré. Il semble donc raisonnable dans ce cas de traiter le bruit de fond de manière non paramétrique. L'objet de ce paragraphe est de proposer une modélisation du bruit de fond par des combinaisons de B-splines. Le lissage proposé dépend d'un paramètre de lissage dont la détermination devrait être automatique (fonction des observations).

Nous nous proposons d'abord de rappeler brièvement quelques propriétés des B-splines. Pour plus de détails le lecteur pourra se reporter aux articles [19], [7] ou [8]. D'une manière informelle et descriptive, une fonction B-spline de degré donné  $q$  sur

$\mathbb{R}$  est une fonction polynomiale par morceaux composée de segments polynomiaux de degré  $q$  adjacents et soumis à certaines contraintes. Par exemple, la fonction B-spline de degré 1 représentée dans la figure 1 est composée de deux segments linéaires : un segment de  $t_1 = 0.25$  à  $t_2 = 0.5$ , l'autre de  $t_2$  à  $t_3 = 0.75$ . Les nœuds de cette B-spline sont les points  $t_1$ ,  $t_2$  et  $t_3$ . À gauche de  $t_1$  et à droite de  $t_3$ , cette fonction est identiquement nulle.

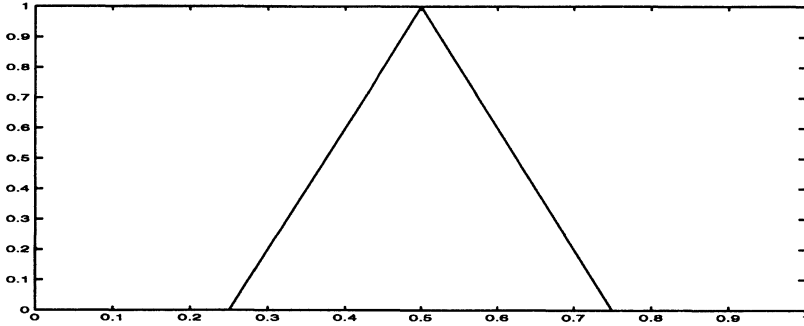


FIGURE 1

*Une fonction B-spline de degré 1 composée de deux segments linéaires :  
un segment de 0.25 à 0.5 et l'autre de 0.5 à 0.75.  
Les nœuds de cette B-spline sont 0.25, 0.5 et 0.75.*

Cet exemple illustre les propriétés générales des B-splines de degré  $q$  :

- ce sont des fonctions polynomiales par morceaux, composées de  $(q+1)$  polynômes de degré  $q$ ;
- ces polynômes se raccordent en  $q$  nœuds intérieurs;
- en chaque nœud intérieur, les dérivées jusqu'à l'ordre  $q-1$  sont continues;
- le support d'une B-spline est restreint à un intervalle délimité par  $q+2$  nœuds. Une B-spline est positive sur son support.
- Tout polynôme d'ordre  $q$  s'écrit comme combinaison linéaire finie de B-splines d'ordre  $q$ .

Cette dernière propriété explique les propriétés d'approximation de fonctions lisses par des bases de B-spline.

Dans l'ouvrage [8] De Boor définit un algorithme pour calculer de manière récurrente une B-spline de degré quelconque à partir de B-splines de degré inférieur et ceci pour un placement arbitraire des nœuds. Si les nœuds sont équirépartis cet algorithme se simplifie considérablement et nous utiliserons cette version dans la suite.

Lorsque le bruit de fond  $B(t)$  est modélisé par une combinaison linéaire de B-splines, l'intensité du signal de diffraction doublement poissonnien s'écrit

$$\mu_t = \sum_{j=1}^m \beta_j B_j(t) + \sum_{k=1}^p c_k f_k(t, \theta_k),$$

où  $m$  est le nombre de nœuds et  $B_j$  sont les fonctions B-splines cubiques de base associées à l'ensemble des nœuds. Avec les notations du paragraphe 4, le vecteur  $n$ -dimensionnel  $\boldsymbol{\eta}$  modélisant l'espérance du vecteur des observations s'écrit

$$\boldsymbol{\eta}(\boldsymbol{\beta}, \boldsymbol{\theta}) = B_{nm}\boldsymbol{\beta} + \boldsymbol{\gamma}(\boldsymbol{\theta})$$

où  $B_{nm}$  est la matrice d'ordre  $n \times m$  dont l'élément  $(i, j)$  est donné par  $B_j(t_i)$  et où  $\boldsymbol{\gamma}(\boldsymbol{\theta})$  est le vecteur  $n$ -dimensionnel représentant l'échantillonnage en les points  $t_i$  de la partie du signal associée aux pics. Lorsque le nombre  $m$  de nœuds retenus est relativement grand, le signal ajusté présentera de fortes variations (faible biais mais grande variance). Afin de réduire cette variance, comme il est classique en régression non paramétrique, on introduira une pénalisation sur l'ajustement du bruit de fond par maximum de vraisemblance. La fonction de pénalisation que nous nous proposons d'utiliser est inspirée des travaux de Eilers [11] et de Eilers et Marx [12] sur les problèmes de régression non paramétrique des moindres carrés par P-splines. Le paramètre  $\lambda$  de sur ou sous-dispersion n'étant qu'un paramètre fantôme pour l'estimation du signal, et afin d'alléger les notations, nous n'aborderons ici que le cas de  $\lambda = 1$ . Nous maximiserons donc la fonction log de vraisemblance pénalisée suivante :

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{t=1}^n y_t \log(\eta_t) - \eta_t - \frac{\alpha}{2} \sum_{j=3}^m (\Delta^2 \beta_j)^2,$$

où  $\Delta$  est l'opérateur de différence défini par  $(\Delta u)_k = u_k - u_{k-1}$  et  $\Delta^2 \beta_j = (\Delta \Delta \beta)_j$ . Cette pénalisation sur le carré de la différence d'ordre 2 entre deux coefficients B-spline contigus est étroitement liée à la pénalisation plus classique sur l'intégrale du carré de la dérivée seconde du signal en régression par splines naturelles (voir [18]).

Avec ces notations on obtient :

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = {}^t \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} + \frac{\partial}{\partial \boldsymbol{\beta}} \left( -\frac{\alpha}{2} \sum_{j=3}^m (\Delta^2 \beta_j)^2 \right).$$

En notant  $D_2$  la représentation matricielle de l'opérateur  $\Delta^2$  et en désignant par  $\mathbf{u}(\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\theta}))$  le vecteur score  $\partial \mathcal{L} / \partial \boldsymbol{\mu}$  de composantes  $(y_i / \mu_i - 1)$ , on a finalement :

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = {}^t B \mathbf{u}(\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\theta})) - \alpha {}^t D_2 D_2 \boldsymbol{\beta}.$$

Notons que la matrice  $D_2$  est la matrice d'ordre  $m - 2 \times m$ , à bande, donnée par :

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}.$$

D'autre part, on a :

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{t \partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = {}^t X(\boldsymbol{\theta}) \mathbf{u}(\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\theta}))$$

où  $X(\boldsymbol{\theta}) = \partial \boldsymbol{\mu} / \partial \boldsymbol{\theta}$ .

Pour les matrices des dérivées secondes, en posant

$$J_n(\boldsymbol{\mu}) = - \frac{\partial \mathbf{u}(\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\mu}} = - \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\mu}^t \partial \boldsymbol{\mu}} = \text{Diag}(y_i / \mu_i^2),$$

on trouve :

$$\begin{aligned} - \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} &= {}^t B J_n(\boldsymbol{\mu}) B + \alpha^t D_2 D_2, \\ - \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^t \partial \boldsymbol{\theta}} &= {}^t X(\boldsymbol{\theta}) J_n(\boldsymbol{\mu}) X(\boldsymbol{\theta}), \\ - \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^t \partial \boldsymbol{\beta}} &= {}^t B J_n(\boldsymbol{\mu}) X(\boldsymbol{\theta}). \end{aligned}$$

L'algorithme des scores des paragraphes précédents devient, avec  $I(\boldsymbol{\mu}) = \mathbb{E}(J_n(\boldsymbol{\mu})) = \text{Diag}(1/\mu_i)$  :

$$\begin{bmatrix} \boldsymbol{\theta}^{r+1} \\ \boldsymbol{\beta}^{r+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}^r \\ \boldsymbol{\beta}^r \end{bmatrix} + \alpha_r \left[ \begin{array}{cc} {}^t X(\boldsymbol{\theta}^r) I(\boldsymbol{\mu}(\boldsymbol{\theta}^r)) X(\boldsymbol{\theta}^r) & {}^t X(\boldsymbol{\theta}^r) I(\boldsymbol{\mu}(\boldsymbol{\theta}^r)) B \\ {}^t B I(\boldsymbol{\mu}(\boldsymbol{\theta}^r)) {}^t X(\boldsymbol{\theta}^r) & {}^t B I(\boldsymbol{\mu}(\boldsymbol{\theta}^r)) B + \alpha^t D_2 D_2 \end{array} \right]^{-1} \mathbf{V},$$

où

$$\mathbf{V} = \begin{bmatrix} {}^t X(\boldsymbol{\theta}^r) \mathbf{u}(\boldsymbol{\mu}(\boldsymbol{\theta}^r, \boldsymbol{\beta}^r)) \\ {}^t B \mathbf{u}(\boldsymbol{\mu}(\boldsymbol{\theta}^r, \boldsymbol{\beta}^r)) - \alpha^t D_2 D_2 \boldsymbol{\beta}^r \end{bmatrix}.$$

L'algorithme précédent détermine les paramètres d'affinement pour une valeur fixée du paramètre de lissage  $\alpha$ . Une valeur de  $\alpha$  égale à 0 conduit à ne pas lisser le bruit de fond et par suite à confondre alors les pics et le bruit de fond. Par contre, une grande valeur de  $\alpha$  conduit à l'ajustement du bruit de fond par un polynôme du second degré et réalise l'affinement obtenu lorsque le bruit de fond est introduit de manière paramétrique. Le choix du paramètre de lissage est important et on cherchera une valeur optimale réalisant le meilleur compromis pour bien estimer simultanément les pics et le bruit de fond. Pour déterminer cette valeur optimale, on a souvent recours au critère de validation croisée (voir [18]). Il s'agit d'un critère du choix du degré de régularité fondé sur la capacité prédictive du modèle ajusté. Dans notre cas, en utilisant comme mesure de qualité de prédiction la déviance, cela revient à minimiser par rapport à  $\alpha$  la quantité :

$$\sum \frac{1}{\hat{\mu}_i} \left( \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2 / (1 - A_{ii})^4$$



où  $\hat{\mu}_i$  est la valeur estimée de  $\mu_i$  pour la valeur courante du paramètre  $\alpha$  et  $A_{ii}$  est le  $i$ -ème terme diagonal de la matrice de lissage associée à la même valeur du paramètre qui est donnée par :

$$B({}^t B J_n(\boldsymbol{\mu}) B + \alpha {}^t D_2 D_2)^{-1} {}^t B J_n(\hat{\boldsymbol{\mu}})$$

L'algorithme de validation croisée défini ci-dessus est bien adapté lorsque l'intensité du processus poissonnien observé ne contient pas de pic. Lorsque des pics sont présents, des simulations ont montré (voir par exemple la figure 2) que l'ajustement global du signal est correct mais ne lisse pas suffisamment le bruit de fond entraînant ainsi un biais important sur l'intensité des pics.

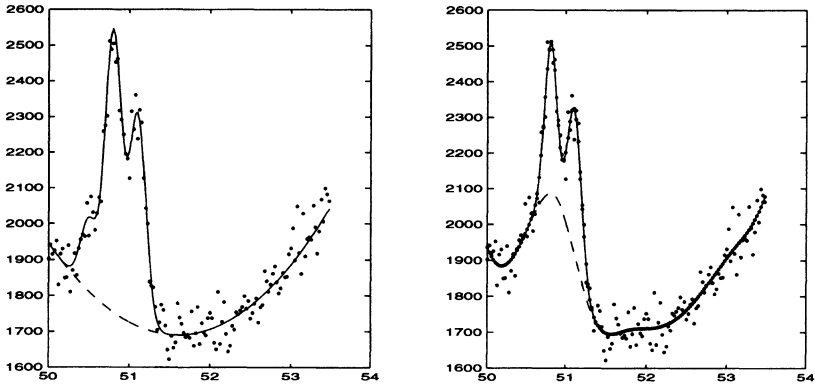


FIGURE 2

*A gauche on observe l'intensité (courbe en trait plein) ayant servi à générer les données représentées par des points ainsi que le bruit de fond tracé en pointillé (polynôme du second degré). A droite l'intensité globale estimée par validation croisée ( $\alpha = 4 \cdot 10^{-10}$ ) est raisonnable alors que l'estimation du bruit de fond, et par suite, celle des pics est fortement biaisée.*

Le mauvais comportement de l'algorithme de validation croisée est dû au fait que ce dernier traite l'intensité  $\mu$  de façon entièrement non paramétrique réalisant ainsi un ajustement global correct sans tenir compte de la structure du signal (bruit de fond auquel sont superposés des pics). Pour résoudre ce problème, supposons momentanément que les fonctions paramétriques modélisant les pics sont connues. Dans ce cas, le bruit de fond s'écrit :

$$\nu_i = \mu_i - f(\theta, t_i) = g(t_i) = (B\boldsymbol{\beta})_i$$

et on peut associer aux observations les pseudo-observations  $z_i = y_i - f(\theta, t_i)$  où  $f(\theta, t_i)$  est la contribution des pics au signal. La matrice d'information de l'algorithme des scores avec lissage s'écrit :

$$J_{bis}(\boldsymbol{\beta}) = \text{Diag} \left( \frac{1}{\nu_i} \right).$$

L'algorithme de validation croisée associé à l'algorithme des scores avec lissage consisterait à choisir  $\alpha$  minimisant

$$\sum \frac{1}{\nu_i} \left( \frac{z_i - \nu_i}{\nu_i} \right)^2$$

Nous utiliserons cet algorithme en remplaçant dans l'expression des pseudo-observations  $z_i$  le vecteur  $\theta$  des paramètres inconnus par son estimation  $\hat{\theta}$  obtenue à la dernière itération de l'algorithme des scores. Les exemples simulés et les applications développées dans le paragraphe suivant montrent que ce critère est raisonnable. Dans le contexte de l'exemple illustré par la figure 2 on aboutit à l'affinement de la figure 3.

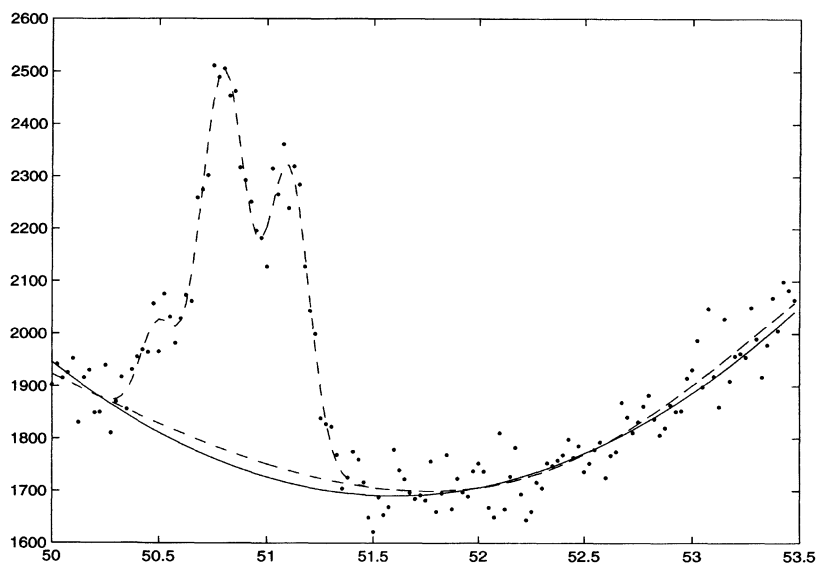


FIGURE 3

*L'intensité estimée (courbe en pointillé) ainsi que l'estimation du bruit de fond (courbe en pointillé) et par suite celle des pics sont obtenues par validation croisée modifiée ( $\alpha = 8 \cdot 10^{-4}$ ). La courbe en trait plein représente le bruit de fond ayant servi à la simulation.*

## 6. Exemples simulés et applications

Afin de valider empiriquement la méthode d'estimation développée au cours des paragraphes précédents nous avons considéré une forme de spectres que l'on retrouve fréquemment dans les modélisations des diagrammes de diffraction de poudre. Plus

précisément les intensités considérées dans l'étude empirique qui suit sont :

$$\begin{aligned}
 f_1(x) &= 1692.125 + 30(x - 51.75) + 100(x - 51.75)^2 \\
 &\quad + 50 \cdot \frac{1}{\sqrt{2\pi \cdot 0.0104}} \exp\left(-\frac{1}{2} \frac{(x - 50.5)^2}{0.0104}\right) \\
 &\quad + 200 \cdot \frac{1}{\sqrt{2\pi \cdot 0.0104}} \exp\left(-\frac{1}{2} \frac{(x - 50.8)^2}{0.0104}\right) \\
 &\quad + 150 \cdot \frac{1}{\sqrt{2\pi \cdot 0.0104}} \exp\left(-\frac{1}{2} \frac{(x - 51.1)^2}{0.0104}\right) \\
 f_2(x) &= 1692.125 + 30(x - 75) + 100(x - 75)^2 \\
 &\quad + 73.45 \cdot \frac{1}{\sqrt{2\pi \cdot 0.052}} \exp\left(-\frac{1}{2} \frac{(x - 50.8)^2}{0.052}\right) \\
 &\quad + 58.22 \cdot \frac{1}{\sqrt{2\pi \cdot 0.056}} \exp\left(-\frac{1}{2} \frac{(x - 51.12)^2}{0.056}\right).
 \end{aligned}$$

Ces deux fonctions ayant servi aux simulations sont représentées sur la figure 4.

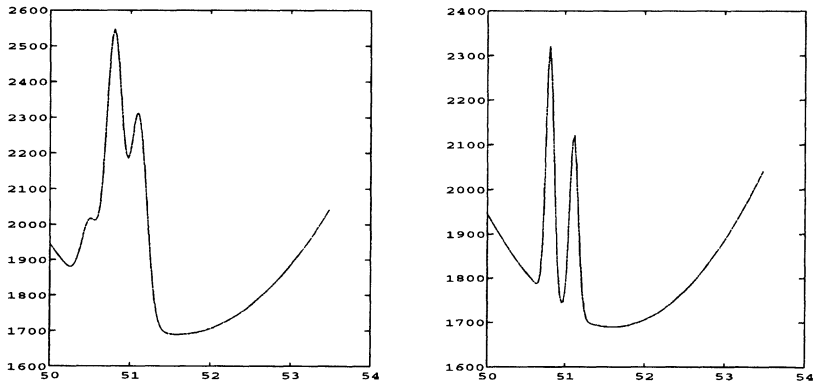


FIGURE 4

Les deux spectres  $f_1$  (à gauche) et  $f_2$  (à droite) ayant servi aux simulations.

Pour chacune de ces fonctions nous avons simulé 100 fois un processus de comptage d'intensité moyenne le spectre en question discrétisée sur 150 points. Le processus de comptage à été simulé soit selon une loi de Poisson, soit selon une loi doublement exponentielle avec un paramètre de dispersion égal à 2 ou à 0.5. Pour chacun des diagrammes simulés de cette manière, les paramètres de l'intensité moyenne et le paramètre de dispersion ont été estimés par la méthode des scores des paragraphes précédents. Pour chaque modèle ceci donne donc une série empirique de 100 valeurs estimées pour chaque paramètre. Les calculs ont été réalisés avec le programme ABFfit [4]. Pour ne pas alourdir la présentation nous ne reportons ici que

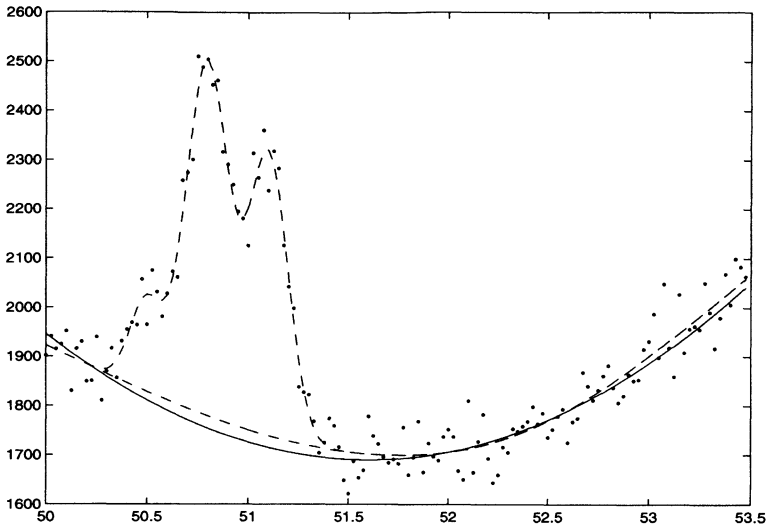


FIGURE 5

*L'intensité estimée (courbe en pointillé) ainsi que l'estimation du bruit de fond (courbe en pointillé) et par suite celle des pics sont obtenues par validation croisée modifiée ( $\alpha = 8 \cdot 10^{-4}$ ). La courbe en trait plein représente le bruit de fond ayant servi à la simulation.*

les résultats des paramètres estimés pour le second pic de la fonction  $f_1$ , ce dernier étant le plus délicat à estimer car partiellement recouvert par les autres pics.

La comparaison (cf. Tableau 1) entre SSAS et SSAA montre que la prise en compte du paramètre de dispersion dans l'affinement n'affecte pas la qualité des estimateurs des autres paramètres. On observe aussi que la variabilité des estimations d'un paramètre est du même ordre de grandeur que l'estimation moyenne de son écart-type. Par contre, SAAS montre que ce n'est plus le cas lorsque le modèle simulé comporte un paramètre de dispersion que l'on n'ajuste pas. Enfin pour SAAA, l'introduction dans l'ajustement du paramètre de dispersion permet de retrouver un bon accord entre l'écart-type estimé des paramètres et leur variabilité dans les simulations.

Pour quantifier l'influence du paramètre de dispersion et valider la méthode d'estimation de ce dernier, les données ont été simulées avec les dispersions 2, 1 et 0.5. Les résultats sont reportés dans le tableau 2. On constate les bonnes qualités de la méthode d'estimation.

Enfin pour illustrer le comportement de l'algorithme de lissage du bruit par B-splines et l'influence éventuelle de ce lissage sur les paramètres des pics nous avons procédé à 100 simulations d'un modèle possédant une intensité égale à  $f_2$ . Le modèle ajusté de manière semi-paramétrique avec la méthode d'estimation du paragraphe 5 conduit aux résultats du tableau 3. On constate, comme on l'espérait, que les estimations des pics sont pratiquement sans biais.

TABLEAU 1

Résultats des estimations des paramètres du second pic de la fonction  $f_1$ .  
 Signification des acronymes : SSAS = Simulation Sans dispersion  
 et Ajustement Sans tenir compte d'une dispersion,  
 SSAA = Simulation Sans dispersion et Ajustement Avec une éventuelle dispersion,  
 SAAS = Simulation Avec dispersion (0.5)  
 et Ajustement Sans dispersion et enfin SAAA = Simulation Avec dispersion (0.5)  
 et Ajustement Avec une éventuelle dispersion. Pour chaque paramètre  
 et pour chaque modèle on retrouve la moyenne du paramètre estimé  
 sur les 100 simulations (avec entre parenthèses son écart-type empirique)  
 ainsi que la variance estimée du paramètre (et son écart-type)  
 par la méthode des scores.

	SSAS	SSAA	SAAS	SAAA
Intégrale du pic (200)	194.1 (25)	194.1 (25)	196.5 (16)	196.5 (16)
Var. Est.	20 (10)	21 (10)	20 (7)	14 (5)
Position (50.8)	50.8005 (0.0067)	50.8005 (0.0067)	50.8007 (0.0047)	50.8007 (0.0047)
Var. Est.	0.0065 (0.0013)	0.0065 (0.0014)	0.0065 (0.0009)	0.0046 (0.0008)
Largeur (0.24)	0.237 (0.023)	0.237 (0.023)	0.238 (0.016)	0.238 (0.016)
Var. Est.	0.021 (0.004)	0.021 (0.004)	0.022 (0.003)	0.016 (0.002)

TABLEAU 2

Valeurs et variances estimées du paramètre de dispersion sur 100 simulations  
 pour des données similaires à celles du tableau 1.

	$\lambda = 2$	$\lambda = 1$	$\lambda = 0.5$
$\hat{\lambda}$	1.97 (0.487)	0.992 (0.201)	0.495 (0.098)
Var. Est.	0.42 (0.104)	0.211 (0.043)	0.106 (0.021)

TABLEAU 3

Résultats des simulations pour un modèle ajusté  
 de manière semi-paramétrique.

	Amplitude	Position	Largeur
Pic 1			
Vrai	73.4542	50.7968	0.0027
Estimé	72.8078	50.7963	0.0026
Écart-type	4.8805	0.0033	0.0003
Pic 2			
Vrai	58.2254	51.1178	0.0032
Estimé	58.1758	51.118	0.0032
Écart-type	4.8394	0.0047	0.0005

### 6.1. Exemples réels

L'exemple de la figure 6 correspond à un diagramme de poudre du composé  $\text{Ho}_8\text{Mg}_{42}\text{Zn}_{50}$ , un approximant de quasicristaux, enregistré sur le diffractomètre à neutrons D1B de l'Institut Laue Langevin (ILL) à Grenoble [16]. L'échantillon étudié comporte une partie cristalline responsable des pics et une partie amorphe qui donne à la ligne de base des pics une ondulation de grande amplitude. Dans la présente étude cette ligne de base est le bruit de fond car l'information recherchée se trouve dans les pics. Les pics sont initialisés de façon raisonnable, le bruit de fond est volontairement initialisé loin de la solution optimale pour mieux illustrer la convergence de l'affinement.

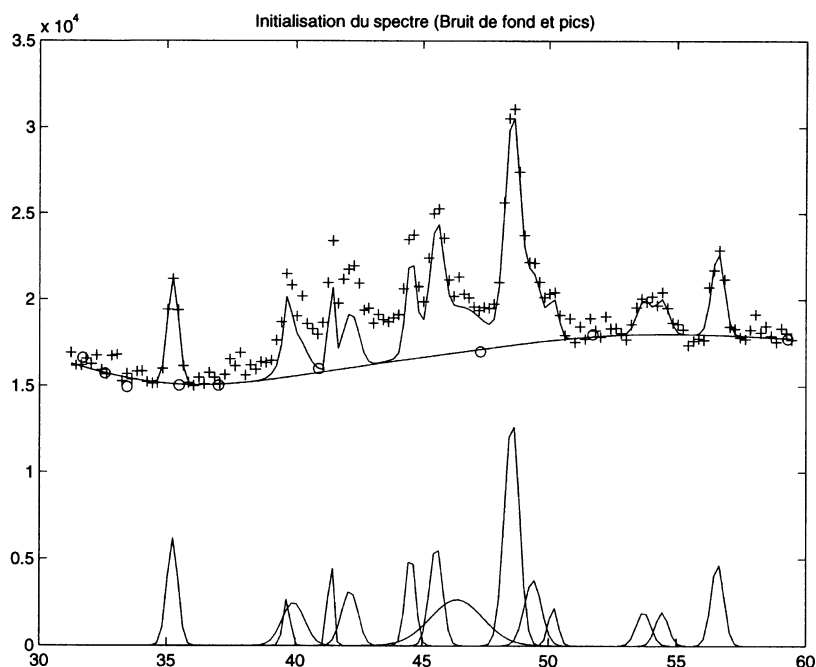


FIGURE 6

*La partie haute de la figure donne les 140 observations (croix), le spectre calculé initial et le bruit de fond (lignes continues) et enfin la position des nœuds des B-splines cubiques (cercles). La partie basse donne la contribution des pics à l'initialisation.*

La figure 7 correspond à la convergence de l'algorithme semi-paramétrique vers une solution satisfaisante au bout de 7 itérations et pour une valeur du paramètre de lissage, optimisée par validation croisée, et égale à  $8 \cdot 10^{-6}$ .

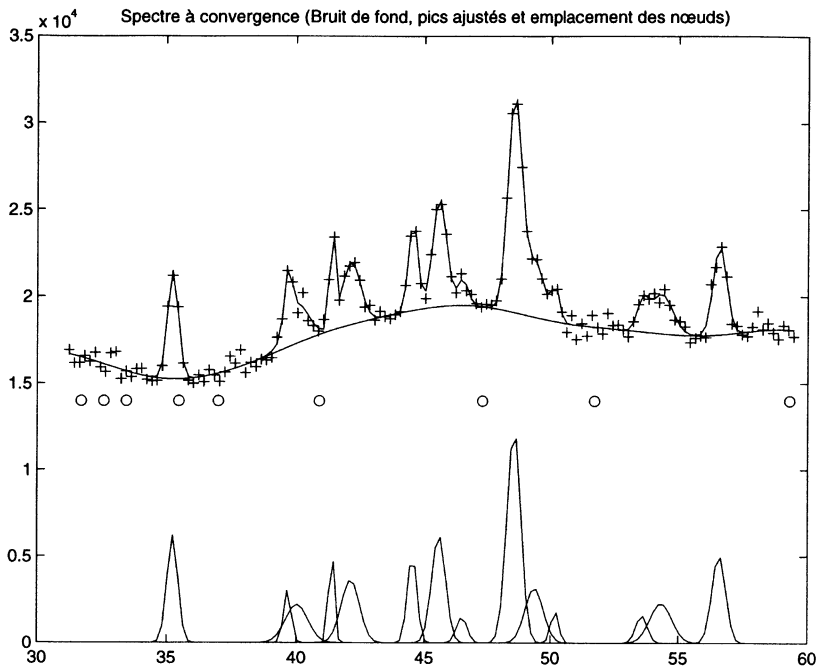


FIGURE 7

*La partie haute de la figure donne les observations (croix),  
le spectre et le bruit de fond ajustés (lignes continues).  
La partie basse donne la contribution des pics estimés à la convergence.  
Les cercles rappellent la positions des nœuds.*

Le bruit de fond de ce diagramme expérimental présente une ondulation qui ne peut être correctement modélisée par un polynôme de faible degré. La stratégie la plus courante dans ce cas pour un cristallographe est, soit d'imposer un bruit de fond a priori (courbe interpolée sur quelques points), soit d'introduire un ou plusieurs pics fictifs très larges dont le seul rôle est de prendre en compte l'ondulation aussi bien que possible. Dans le premier cas, toute erreur dans la forme arbitraire du bruit biaise l'estimation des paramètres des pics. Dans le second cas les résultats dépendent largement des hypothèses faites sur la forme du bruit de fond. L'approche semi-paramétrique repose sur le fait que le physicien a plus de renseignements sur les pics que sur le bruit de fond car il dispose souvent d'un modèle physique permettant de décrire la partie «pics» du diagramme. Dans l'approche semi-paramétrique le bruit de fond estimé est influencé par l'affinement des pics alors que dans l'approche paramétrique c'est le bruit de fond qui influence l'estimation des paramètres des pics. Pour donner une souplesse suffisante à la modélisation du bruit de fond nous avons utilisé, dans cet exemple, une combinaison de 9 B-splines avec des nœuds non régulièrement répartis. En effet nous avons constaté qu'avec un trop petit nombre de nœuds, le lissage agissant de façon globale, l'algorithme présente une grande sensibilité aux valeurs d'initialisation et, par conséquent, une perte de robustesse.

*Remerciements*

Les auteurs remercient les deux rapporteurs et l'éditeur pour leur lecture méticuleuse de la version manuscrite, ainsi que pour leurs critiques et leurs suggestions qui ont permis d'améliorer la présentation de ce travail.

**Références**

- [1] ABRAMOVITZ M., STEGUN I.A. (1965). *Handbook of Mathematical Functions*. Dover, New York.
- [2] AITKIN M. (1995). Probability model choice in single samples from exponential families using Poisson log-linear modeling, and model comparison using Bayes and posterior Bayes factors. *Statistics and Computing*, **5**, 113–120.
- [3] ANTONIADIS A., BERRUYER J., CARMONA R.(1992). *Régression non linéaire et applications*. Economica, Paris.
- [4] ANTONIADIS A., BERRUYER J., FILHOL A. (1990). Maximum likelihood estimation methods in powder diffraction refinements. *Acta. Cryst.*, A46, pp. 692–711.
- [5] BARNDORFF-NIELSEN O.E. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons, New York.
- [6] BARNDORFF-NIELSEN O.E., COX D.R.(1989). *Asymptotic techniques for use in statistics*. Chapman & Hall, New York.
- [7] BESSE Ph., THOMAS-AGNAN Ch. (1989). Le lissage par fonction splines en statistique : revue bibliographique. *Statistique et Analyse des données*, **14** n° 1, 55–84.
- [8] DE BOOR C.(1978). *A practical guide to splines*. Springer-Verlag, Berlin.
- [9] DIACONIS P., EFRON B. (1985). Testing the independence of a two-way table : New interpretations of chi-square statistics. *Ann. Statist.*, **13**, 845–913.
- [10] EFRON B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.*, **81**, 709–721.
- [11] EILERS P.H.C.(1990). Smoothing and interpolation with generalized linear models. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali*, **12**, 21–32.
- [12] EILERS P.H.C. , MARX B.D.(1996). Flexible smoothing with *B*-splines and penalties. *Statistical Sciences*, **11**, 89–121.
- [13] JØRGENSEN B. (1987). Exponential dispersion models. *J.R.S.S., Ser B*, **49**, 127–162.
- [14] NELDER J.A., WEDDELBURN R.W.M. (1972). Generalized linear interactive models. *J.R.S.S., Ser A*, **135**, 370–384.
- [15] RAO C.R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New York.
- [16] SCHMITT F., CHARIER B., OULADIAFF B.(1998). Communication privée.



- [17] SEEBER G.U.H. (1997). Overdispersed exponential regression models. *Computational Statistics*, **12**, 209–218.
- [18] SILVERMAN B.W. (1986). Some aspects of the smoothing spline approach to nonparametric regression curve fitting. *J.R.S.S., Ser B*, **47**, 1–52.
- [19] WEGMAN E.J., Wright I.W. (1983). Splines in Statistics. *J. Amer. Statist. Assoc.*, **78**, 351–365.