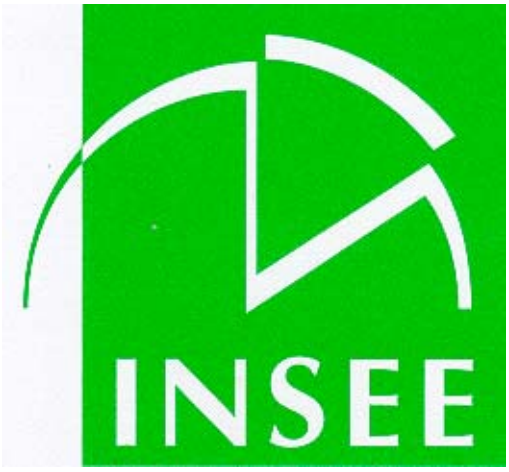
Direction des Statistiques Démographiques et Sociales

N° F0801

APPROCHE MULTIVARIÉE DE L'ESTIMATION DES INEGALITÉS DANS L'ENQUÊTE PATRIMOINE 2004

Eric Gautier
Cédric Houdré

Document de travail



Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des Documents de Travail
de la
DIRECTION DES STATISTIQUES DÉMOGRAPHIQUES ET SOCIALES

Département des Prix à la Consommation,
des Ressources et des Conditions de vie des Ménages

N°F0801

Approche multivariée de l'estimation des inégalités dans l'enquête Patrimoine 2004

Eric GAUTIER
(Ensae-Crest)

Cédric HOUDRÉ
(Insee-Crest, Division Revenus et patrimoine des ménages)

février 2008

Approche multivariée de l'estimation des inégalités dans l'enquête Patrimoine 2004

Résumé :

Dans les enquêtes auprès des ménages, les questions sur les montants d'actifs proposent souvent au ménage de fournir un intervalle plutôt qu'un montant précis. Cette stratégie réduit le taux de non-réponse. En contrepartie, les données de montants sont tantôt en clair, tantôt en intervalles ou complètement manquantes.La production d'indicateurs d'inégalités nécessitent toutefois de travailler sur des données complètes ou complétées par imputation.

Nous présentons et comparons dans cet article trois procédures d'estimation d'indicateurs d'inégalités de patrimoine fondées sur de techniques de simulation et appliquées à l'enquête Patrimoine 2004. Les procédures sont des extensions de la méthode des "résidus simulés" étudiée par Lollivier et Verger (1989). Elles se distinguent par la quantité d'information utilisée pour reconstituer des variables de montants continues: d'une information relativement fruste sur le patrimoine total du fait d'une censure importante pour les patrimoines élevés, à une information plus détaillée sur les composantes de ce patrimoine total moblisant également des informations agrégées sur des sommes de composantes ou sur l'imposition du ménage à l'Impôt de Solidarité sur la Fortune.

Nous proposons une méthode hiérarchique qui fournit une estimation par prévision et des intervalles de confiance dont la couverture tient compte des aléas liés au sondage et à la non-réponse totale, ainsi qu'à la simulation. La troisième approche est fondée sur une modélisation jointe des composantes du patrimoine du ménage. Le modèle rend alors compte de la corrélation entre les montants que les ménages allouent à leurs différents actifs patrimoniaux, en dehors de la corrélation liée à l'hétérogénéité observée ou à l'existence de contraintes d'allocation du patrimoine dans un portefeuille fini. L'hétérogénéité inobservée pourrait par exemple contribuer à cette corrélation. Dans ce cas, nous adoptons le point de vue de la statistique Bayésienne et la couverture des intervalles de confiance tient également compte de l'incertitude sur les paramètres du modèle. La simulation repose alors sur un algorithme d'échantillonage de Gibbs.

Mots-clés: patrimoine des ménages, imputation jointe, données censurées, estimation bayésienne, échantillonnage de Gibbs

A multivariate setting for the estimation of wealth ineaquality indicators applied to Enquête Patrimoine 2004

Abstract:

Survey questions that ask respondents to report financial amounts, particulary assets amounts in household wealth surveys, are subject to high rates of item missing data. A strategy commonly used by survey designers to curb this item non-response rate consist in allowing for bracketed response questions. In return, the resulting measures are a mixture of single valued responses, pre-defined brackets responses, self-defined interval responses and completely missing data. The coarsened nature of these data poses special challenges when analysing wealth inequalities for the estimation of inequality indicators, such as Gini coefficients, requires to work on complete data or data completed by way of imputation techniques.

This article presents three procedures of estimation of wealth inequality indicators based on different imputation strategies applied to the data the French household wealth survey "Enquête Patrimoine 2004". The imputation techniques all draw on Lollivier and Verger (1989) adaptation of simulated residuals methods, but the estimation prodecures differ in the nature and quantity of information they rely on: the first estimate is based on an aggregate measure of gross wealth collected via a single question and affected by substantial interval censoring in the upper part of the distribution, the second and third estimates are based on coarsened information on more detailed wealth components as well as on additional information whether the household is subject to the wealth tax.

The hierarchical approach of the estimation followed gives interval confidence estimates accounting for sampling and total non-response error, as well as simulation error for completed data. The third estimation procedure rely on joint imputation of several wealth components. The model accounts for the potential correlation between the amounts allocated to different assets, even after controlling for observed heterogeneity or aggregate constraints on gross wealth or taxable wealth threshold. Unobserved heterogenity could contribute to this correlation. The joint imputation is achieved in a Bayesian setting, which allow as a by product to produce confidence interval accounting for parameters estimation variance. Simulation in the joint law of components is achieved through Gibbs sampling.

Keywords: household wealth, multivariate imputation, coarsened data, bracketed response, Bayesian estimation, Gibbs sampling

Approche multivariée de l'estimation des inégalités dans l'enquête Patrimoine 2004

Eric Gautier[1]
Cédric Houdré[2]

Résumé

Dans les enquêtes sur le patrimoine des ménages, les questions de montants, celles où le ménage doit par exemple donner un montant détenu sur tel ou tel produit financier ou immobilier, proposent souvent au ménage de fournir un intervalle plutôt qu'un montant précis. Cette stratégie permet de réduire le taux de non-réponse. En contrepartie, les données de montants seront tantôt en clair, tantôt en intervalles, tantôt complètement manquantes. Pourtant la production d'indicateurs d'inégalités nécessite a priori de travailler sur des données complètes ou éventuellement complétées par imputation.
Nous présentons et comparons dans cet article trois procédures d'estimation d'indicateurs d'inégalités de patrimoine fondées sur des techniques de simulation et appliquées à l'enquête Patrimoine 2004. Les procédures sont des extensions de la méthode des "résidus simulés" étudiée par Lollivier et Verger (1989). Elles se distinguent par la quantité d'information utilisée pour reconstituer des variables de montants continues: d'une information relativement fruste sur le patrimoine total du fait d'une censure importante pour les patrimoines élevés, à une information plus détaillée sur les composantes de ce patrimoine total mobilisant également des informations agrégées sur des sommes de composantes ou sur l'imposition du ménage à l'Impôt de Solidarité sur la Fortune.
Nous proposons une méthode hiérarchique qui fournit une estimation par prévision et des intervalles de confiance dont la couverture tient compte des aléas liés au sondage et à la non-réponse totale, ainsi que de l'aléa du modèle pour les données d'enquête censurées. La troisième approche est fondée sur une modélisation jointe des composantes du patrimoine du ménage. Le modèle rend alors compte de la corrélation entre les montants que les ménages allouent à leurs différents actifs patrimoniaux, en dehors de la corrélation liée à l'hétérogénéité observée ou à l'existence de contraintes d'allocation du patrimoine dans un portefeuille fini. L'hétérogénéité inobservée pourrait par exemple contribuer à cette corrélation. Dans ce cas, nous adoptons le point de vue de la statistique Bayésienne et la couverture des intervalles de confiance tient également compte de l'incertitude sur les paramètres du modèle. La simulation repose alors sur un algorithme d'échantillonnage de Gibbs.

[1] ENSAE - CREST, Timbre J120, 3 avenue Pierre Larousse, 92240 Malakoff, gautier@ensae.fr
Ce travail a été réalisé en grande partie lorsqu'Eric Gautier travaillait à l'Unité Méthodes Statistiques de la Direction des Statistiques Démographiques et Sociales à l'INSEE.
[2] INSEE - CREST, Division Revenus et Patrimoine des Ménages, timbre F350, 17 boulevard Adolphe Pinard, 75675 Paris Cedex, cedric.houdre@insee.fr

Cet article résulte d'un travail réalise collectivement par la Section Patrimoine des Ménages et l'Unité Méthodes Statistiques de l'INSEE. Marie Cordier, Catherine Rougerie et Siabou Diaby ont contribué de manière décisive à la mise en œuvre des méthodes mises en production pour l'enquête Patrimoine 2004. Nous remercions également Luc Arrondel, Céline Bessière, Pascal Chevalier, Sibylle Gollac, Muriel Roger et Daniel Verger pour les discussions enrichissantes au sein du groupe de travail, et pour leur aide précieuse dans la spécification des modèles utilisés pour l'imputation du fichier produite en 2006. La méthode présentée ici est toutefois différente et a été développée en parallèle. Nous remercions aussi Christian Robert pour ses conseils sur les algorithmes MCMC. Nous remercions enfin Alain Trognon qui a discuté une version préliminaire de ce travail ainsi que les participants du séminaire de la Direction des Statistiques Démographiques et Sociales et ceux du séminaire de recherche en économétrie de l'université de Yale où ce travail a été présenté.

L'analyse microéconomique des inégalités de revenus ou de patrimoine s'appuie généralement sur des indicateurs synthétiques nécessitant l'utilisation de montants mesurés de façon continue. Or les enquêtes auprès des ménages sur le patrimoine font face, en France comme à l'étranger, à une difficulté majeure dans l'observation des encours : la qualité des données. La non réponse partielle est d'une part très forte. Juster et Smith (1997) rapportent que dans les enquêtes américaines Health and Retirement Study (HRS) et Aging and HEAlth Dynamics (AHEAD), les taux de non-réponse aux questions de montants peuvent atteindre 20 à 40 %. D'autre part, même lorsque le ménage fournit une réponse en clair, le montant est souvent déclaré avec une marge d'erreur non négligeable. Pour contourner cet obstacle, il est possible de proposer au ménage de donner une réponse en intervalle. Plusieurs stratégies sont envisageables. L'enquête Patrimoine en retient deux suivant les actifs patrimoniaux considérés. Pour la résidence principale par exemple, le ménage est d'abord interrogé sur le montant réel de sa résidence, et, en cas de non-réponse, est invité à donner des bornes inferieures et supérieures, choisies par lui, encadrant la valeur de son bien. Pour les actifs financiers en revanche, le ménage choisit un intervalle prédéfini parmi une grille proposée par l'enquêteur. Ce procédé permet de substituer une partie des données qui seraient manquantes par des données censurées. Les montants peuvent désormais être de nature différente : valeurs continues, censurées par intervalle (éventuellement non majoré) ou complètement manquantes.

L'analyse des inégalités et le calcul d'estimateurs de sondage basés sur le plan pour les grandeurs d'intérêt, comme l'indice de Gini du patrimoine des français par exemple, nécessitent alors de travailler sur un fichier de données « complètes », c'est-à-dire dans lequel les données manquantes ou partiellement manquantes (censurées) ont été imputées. L'*imputation* est couramment utilisée à l'Insee, mais les techniques sont nombreuses. Dans le cadre des enquêtes Patrimoine, la technique utilisée depuis l'enquête de 1992 est appelée « méthode des *résidus simulés* » par Lollivier et Verger (1989) en hommage à l'article de Gouriéroux et al. (1987b) qui traite de l'utilisation de la simulation de résidus dans des modèles à variables latentes (Probit, Tobit…) afin de construire des représentations graphiques permettant de détecter des variables omises, des valeurs aberrantes ou l'hétéroscédasticité. Les *résidus généralisés* également évoqués par Lollivier et Verger permettent d'effectuer des tests de dépendance sans spécification de modèle joint (Gouriéroux et al. (1987a)). Nous donnons une application de la méthode initiale des résidus généralisés à l'enquête patrimoine dans l'Annexe 2.

La méthode d'imputation étudiée par Lollivier et Verger repose sur une modélisation économétrique des données. Les hypothèses du modèle concernent donc la forme fonctionnelle entre les variables d'intérêt (celles qu'il faut imputer), des variables explicatives observées dans l'enquête et la loi des résidus, qui est éventuellement une loi jointe si on impute simultanément plusieurs composantes. L'idéal pour vérifier la qualité de la méthode consisterait à pouvoir comparer des indicateurs obtenus sur données initialement complètes aux même indicateurs obtenus sur des données rendues incomplètes puis imputées. C'est ce que font Lollivier et Verger à l'aide de données sur le revenu total des français. Dans ce cadre univarié, ils modélisent les données censurées sous forme log-linéaire à l'aide des caractéristiques socio-démographiques observables et supposent que la variable latente de revenu total suit une loi log-normale. Remplacer les données censurées ou manquantes par des données continues simulées dans la loi, tronquée lorsque l'information partielle existe, donne des indicateurs d'inégalité de revenus très voisins de ceux calculés à partir des données initiales non censurées. Cette méthode peut-être appliquée de la même manière sur les données de l'enquête Patrimoine, notamment en utilisant une question récapitulative sur le patrimoine brut total du ménage posée en fin de questionnaire : « si vous aviez à liquider la totalité de ce que votre ménage possède à ce jour […] combien pourriez-vous retirer de la vente ? » Si l'utilisation de cette unique question a évidemment comme avantage principal de pouvoir se limiter à un cadre extrêmement simple, il est néanmoins nécessaire de s'interroger sur la validité des hypothèses de modélisation, ainsi que sur le biais et la précision des estimateurs obtenus pour le Gini des français, ou d'autres indicateurs synthétiques résumant des aspects de la distribution du patrimoine des français, en comparaison des résultats que peuvent fournir d'autres stratégies d'imputation.

L'objectif de l'article est de présenter des estimations de tels indicateurs, mais surtout de discuter des hypothèses de trois stratégies d'imputations : imputation de la variable récapitulative de patrimoine total, imputation dans un cadre univarié, c'est-à-dire, actif par actif, de composantes très détaillées du patrimoine en utilisant des informations récapitulatives de l'enquête mais également de l'information

auxiliaire sur l'imposabilité à l'ISF, et enfin imputation jointe de quelques composantes patrimoniales à l'aide d'une méthode de Monte Carlo par chaîne de Markov : l'échantillonnage de Gibbs, qui plus est dans un cadre Bayésien qui permet de calculer des intervalles de confiance pour les indicateurs tenant compte de l'incertitude sur les paramètres de la modélisation.

Mesurer les inégalités à partir d'une unique variable récapitulative : simplicité de l'information mais sous-utilisation des données

L'avantage d'une question récapitulative sur le patrimoine total comme celle de l'enquête Patrimoine :

> « Si vous aviez à liquider la totalité de ce que votre ménage possède à ce jour […], combien le ménage pourrait-il retirer de la vente ? »

est qu'on dispose de manière directe du niveau de patrimoine total du ménage, encore que ce ne soit qu'un patrimoine brut alors qu'on pourrait souhaiter analyser les inégalités en termes de patrimoine net. Cependant, comme pour beaucoup de montants dans cette enquête, le ménage doit se placer dans un système de tranches. Pour calculer des indicateurs d'inégalité de patrimoine, il est usuel de compléter les données par simulation en s'appuyant sur un modèle pour les données censurées auquel on fera référence sous l'appellation Data Generating Process (DGP). Les intervalles de confiance devront malgré tout tenir compte de l'information imparfaite liée à la censure[3] ainsi que de l'aléa de sondage (Encadré 1).

Encadré 1

Aléa de sondage, vers un modèle hiérarchique

Considérons l'exemple de l'indicateur de Gini relatif à la distribution de patrimoine total de l'ensemble de français. Il se calcule via la formule :

$$G = \frac{\sum_{\mathrm{k}\in\mathrm{U}}(2\mathrm{r(k)} - 1)\mathrm{pt_k}}{N\sum_{\mathrm{k}\in\mathrm{U}}\mathrm{pt_k}} - 1$$

où pt_k est le patrimoine total du ménage indicé par k dans l'ensemble U des ménages français et $\mathrm{r(k)}$ est le rang du patrimoine possédé par le kième ménage. Si les patrimoines sont observés sur un sous-échantillon $\mathrm{s} \subset \mathrm{U}$, un estimateur de sondage basé sur le plan s'obtient par :

$$\hat{G} = \frac{\sum_{\mathrm{k}\in\mathrm{s}}(2\hat{\mathrm{r}}\mathrm{(k)} - 1)\mathrm{w_k pt_k}}{\sum_{k\in U} w_k \sum_{\mathrm{k}\in\mathrm{U}}\mathrm{w_k\ pt_k}} - 1$$

où w_k sont les poids de sondage et $\hat{r}(k) = \sum_{j\in s} w_j 1_{\{pt_j \le pt_k\}}$. Cet estimateur est une grandeur aléatoire du fait que s est tiré au hasard dans U. On donne alors des intervalles de confiance approchés pour G en faisant une approximation normale $\hat{G} \approx G + \sqrt{V(\hat{G})}\epsilon$ où $V(\hat{G})$ est la variance de $\hat{G}$ et ϵ est une variable aléatoire normale centrée réduite et en approchant $V(\hat{G})$ par un estimateur $\hat{V}(\hat{G})$[4]. Ce qui par inversion donne :

$$G \approx \hat{G} + \sqrt{\hat{V}(\hat{G})}\epsilon \quad (I_0).$$

A partir de l'approximation (I) nous pouvons déduire un intervalle de confiance pour G.
Par la suite, souhaitant tenir compte de l'aléa de sondage dans notre inférence nous prendrons comme point de départ cette inversion et ferons comme si G est aléatoire et posons

$$G = \hat{G} + \sqrt{\hat{V}(\hat{G})}\epsilon \quad (I).$$

Etant donné que $\hat{G}$ et $\hat{V}(\hat{G})$ sont alors des "paramètres" inconnus nous les modéliserons eux aussi, ce qui revient à modéliser les données d'enquête censurées. Nous qualifions de modèle hiérarchique cet empilement de deux modèles.

[3] Sous la forme soit d'un aléa de modèle si le modèle est connu, soit d'un aléa de modèle et d'un aléa lié à l'incertitude sur les paramètres à estimer si le modèle appartient à une famille paramétrée.

[4] Pour les calculs de l'estimateur de la variance d'indicateurs d'inégalités nous renvoyons à Dell et al. (2002).

Lollivier et Verger estiment une régression log-linéaire censurée du revenu total. Ils complètent ensuite l'échantillon par simulation (résidu simulé) ou prévision (résidu généralisé) puis calculent des indicateurs d'inégalité sur ces données reconstruites et comparent les résultats obtenus avec les résultats calculés directement sur l'échantillon initial complet. Ils concluent que compléter le fichier par simulation donne des résultats très voisins des résultats sur les vraies données alors que la prévision donne des résultats biaisés. Ce résultat constaté empiriquement s'explique de par la non-linéarité des grandeurs d'intérêt (Gini, quantiles…) en les données manquantes ou censurées (Encadré 2). Dans le cadre de ce modèle, on peut choisir de fournir une unique valeur pour chacune des grandeurs d'intérêt, mais dans ce cas il importe de définir un critère d'optimalité (Encadré 2), ou de donner un intervalle de confiance tenant compte en plus de l'aléa de sondage (Encadré 1) de l'incertitude liée à la censure des données (Encadré 2).

Encadré 2

Inférence par simulation sur données d'enquêtes manquantes ou censurées

Data Generating Process et sélection des ménages

Nous posons pour les patrimoines pt_k de l'échantillon s, de taille m, un modèle conditionnel à s, correspondant à des patrimoines tirés dans une super-population, puis échantillonnés via le sondage, puis par non-réponse totale. Les grandeurs $\hat{G}$ et $\hat{V}(\hat{G})$ apparaissant dans (I) sont désormais aléatoires. Cet aléa est du à l'incertitude sur les pt_k : les intervalles et les observations des covariables X_k ne permettent pas de connaitre exactement pt_k. Nous supposons ici connues les valeurs des paramètres du DGP, l'approche Bayésienne détaillée plus loin permet par exemple de prendre en compte également cette incertitude.
L'introduction de covariables a un intérêt prédictif, minimisant l'aléa de modèle. Bien entendu l'ajout de variables à ses limites car en contrepartie on perd en précision sur la connaissance des paramètres du modèle car nous ne disposons cette fois-ci d'un échantillon fini de taille fixé à l'avance. La couverture d'un intervalle de confiance peut donc être de la sorte améliorée. Les covariables peuvent aussi parfois résoudre des problèmes de sélection exogène (Gautier (2005) et Little et Rubin (2002)). Il est en effet ici important de pouvoir dire, en ayant conditionné par les observables pertinents, que ϵ_i est indépendant de $(pt_1^i, \dots, pt_m^i)$. Nous ferons cette hypothèse d'ignorabilité dans la suite.

Optimalité

Supposons que l'on doive donner une unique valeur G^* pour les grandeurs d'intérêt et que les données censurées soient issues d'un DGP connu. Il s'agit donc d'un problème de prévision. L'optimalité est définie à l'aide d'un critère de risque.
Dans le cas du risque quadratique, le G^* optimal minimise :

$$E\left[\left(G^* - G(pt_1, \dots, pt_m)\right)^2 \middle| X_1 = x_1, pt_1 \in \left[pt_{1,min}, pt_{1,max}\right], \dots, X_m = x_m, pt_m \in \left[pt_{m,min}, pt_{m,max}\right]\right] \quad (R)$$

où $G(pt_1, \dots, pt_m)$ est donné par le modèle hiérarchique composé de (I) et du modèle conditionnel de pt_k sachant X_k (DGP). L'optimal est atteint par l'espérance conditionnelle

$$E\left[G(pt_1, \dots, pt_m) \middle| X_1 = x_1, pt_1 \in \left[pt_{1,min}, pt_{1,max}\right], \dots, X_m = x_m, pt_m \in \left[pt_{m,min}, pt_{m,max}\right]\right].$$

Fournir un résultat issu d'une unique complétion des données de s est donc sous-optimal. Un scenario pourrait être exceptionnellement haut ou bas ce qui nous fait courir un risque élevé de le publier. C'est pourtant l'imputation aléatoire simple qui est utilisée dans Lollivier et Verger (1989). Cette espérance peut être approchée simplement par méthode de Monte-Carlo, c'est-à-dire par sa contrepartie empirique :

$$\frac{1}{T}\sum_{i=1}^{T} G^i\left(pt_1^i, \dots, pt_m^i\right)$$

où les pt_k^i sont simulés dans la loi du DGP tronquée (conditionnelle à l'information $pt_k \in \left[pt_{k,min}, pt_{k,max}\right]$). Enfin, avec ce point de vue il est évident que procéder par prévision des données

censurées ne peut pas avoir de bonnes propriétés ici, en effet $G^i(pt_1^i, \dots, pt_m^i)$ n'est pas linéaire en les pt_k^i, condition sur f pour que $\int f(g(x))dx = f(\int g(x)dx)$.

Intervalles de confiance

Plutôt que de donner une unique valeur il peut être souhaitable de fournir un intervalle de confiance des grandeurs d'intérêt.

A partir de réalisations de $(pt_1^i, \dots, pt_m^i)$ dans la loi conditionnelle aux observations des X_k et aux tranches (lois tronquées) et de ϵ_i (modèle *(I)*) nous disposons de scénarii G^i pour le Gini des français. Si ces scénarii sont indépendants, la loi des grands nombres nous donne que les quantiles empiriques convergent vers les vrais quantiles et nous pouvons donc obtenir des intervalles de confiance pour G :

$$\lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{T} 1_{\{G^i(pt_1^i,\dots,pt_m^i)\in I^c\}} = \mathbb{P}\left[G^i(pt_1^i, \dots, pt_m^i) \in I^c \middle| X_1 = x_1, pt_1 \in [pt_{1,min}, pt_{1,max}], \dots, X_m = x_m, pt_m \in [pt_{m,min}, pt_{m,max}]\right] p.s.$$

Le paramètre T est choisit par le statisticien, il peut être pris très grand de sorte que cette approximation soit de très bonne qualité. Les intervalles de confiance sont obtenus en inversant cette fonctionnelle (nous les prenons ici symétriques).

Sur les données de l'enquête Patrimoine, le DGP pour le patrimoine total reprend l'hypothèse de log-normalité des résidus faite par Lollivier et Verger. L'enquête fournit des régresseurs tels que l'âge, le niveau de revenus, la composition familiale, ou encore le statut d'activité (salarié / indépendant) et l'existence d'héritage ou de donations reçus. L'estimation des paramètres du modèle permet de calculer une prédiction du patrimoine, fondée sur les caractéristiques observables du ménage, et de simuler un résidu dans la loi normale conditionnelle à la tranche déclarée.

Compte-tenu de l'importance du patrimoine immobilier dans le patrimoine total des français, nous avons distingué la sous-population des ménages propriétaires de celle des ménages non propriétaires. Les régresseurs comprennent des caractéristiques économiques et socio-démographiques des ménages, ainsi que des indicatrices de la composition du patrimoine en détention. Avec cette méthode, l'indice de Gini serait compris entre 0.60 et 0.63 à 95% (Tableau 1).

Tableau 1 : Indicateurs d'inégalités à partir d'un modèle univarié sur la variable récapitulative

Grandeur d'intérêt	Prévision	Intervalle de confiance	
		Borne inférieure	Borne supérieure
Moyenne	186 366,89	179 622,82	192 994,48
Médiane	113 023,48	107 966,15	117 919,86
P99	1 370 161,66	1 218 001,84	1 535 631,61
P95	605 770,32	572 422,94	637 159,27
P90	413 240,45	397 980,20	427 342,64
Q3	230 369,85	223 560,28	236 644,55
Q1	18 324,53	16 930,28	19 789,70
P10	4 809,65	3 958,76	5 631,70
P95/D5	5,3632	5,0157	5,6915
P99/D5	12,1303	10,7528	13,6488
Q3/Q1	12,5922	11,6059	13,5550
D9/D1	86,2650	70,9916	101,1864
D9/D5	3,6587	3,4754	3,8278
Gini	0,6171	0,6035	0,6306
Theil	0.7475	0,7046	0,7946
Atkinson ($\epsilon = 1.5$)	0,8420	0,8341	0,8497
Atkinson ($\epsilon = 2$)	0,9251	0,9159	0,9339

Source : Enquête Patrimoine 2004, $T = 1000$, calculs des auteurs.

Lecture : Estimateur ponctuel et bornes de l'intervalle de confiance à 5%. L'intervalle de confiance ne tient pas compte de l'incertitude sur les paramètres liés à la phase d'estimation.

Cependant, la largeur des intervalles de confiance présentés dans le Tableau 1 est sous-évaluée. Il est important de tenir compte de l'incertitude liée à la phase d'estimation des paramètres, ce que nous ferons plus facilement dans le cadre Bayésien que nous présentons dans la suite (Encadré 3).
D'autres problèmes se posent avec cette méthode.

L'hypothèse de log-normalité : une hypothèse non testable en l'absence de données complètes

Une des hypothèses fondamentales de la simulation est le choix d'une forme pour la distribution du patrimoine total des français. Les lois log-normales sont souvent bien adaptées pour décrire la distribution des revenus ou celle des salaires. On peut objecter qu'une loi de Pareto pourrait être mieux adaptée pour décrire celle des patrimoines[5]. En tout état de cause, il est impossible de tester cette hypothèse en l'absence de données complètes au niveau microéconomique sur le patrimoine des français. Il existe certes des données fiscales sur le patrimoine mais 2% des ménages seulement sont concernés par l'impôt sur la fortune et le concept de patrimoine imposable ne recouvre que partiellement celui mesuré dans l'enquête Patrimoine.
Cette hypothèse pose par ailleurs un autre problème lié à la sélection de notre échantillon. En effet, compte tenu de la très forte concentration du patrimoine, l'échantillon de l'enquête 2004 surreprésente certaines catégories de la population comme les indépendants, les cadres, les retraités ainsi que les quartiers « aisés »[6] de manière à améliorer le biais et la précision des estimateurs de sondage basés sur le plan. La sélection est donc volontairement liée à la variable d'intérêt et nous avons veillé à la rendre exogène dans l'estimation en incluant les variables ayant servi à cette surreprésentation parmi les régresseurs quand elles étaient significatives. Tenir compte de ces variables est essentiel car s'il est crédible que la loi du patrimoine des français soit log-normale, la loi du patrimoine des ménages sélectionnés de façon endogène, n'a que peu de chances d'être elle aussi log-normale.

La censure des patrimoines élevés conduit à des estimateurs potentiellement biaisés et moins précis

Reconstituer le patrimoine total à partir de la variable récapitulative pose un autre défi. Le seuil de la tranche la plus haute s'élève « seulement » à 450 000 €. Les ménages possédant un patrimoine plus élevé contribuent très peu à la vraisemblance des modèles, qui sont par conséquent valables essentiellement pour les ménages situés en dessous de ce seuil. La censure peut alors poser problème dans le cas où le modèle est mal spécifié. L'homogénéité des coefficients de régression entre les ménages situés au milieu et en haut de la distribution du patrimoine peut, en l'absence d'information suffisante sur le haut de la distribution conduire à un biais. L'homoscédasticité des résidus également : il se pourrait que le patrimoine soit plus dispersé pour les ménages à hauts revenus par exemple. Ne pas pouvoir prendre en compte cette hétéroscédasticité par manque d'information sur les patrimoines élevés conduirait alors à des patrimoines simulés vraisemblablement trop bas pour les ménages « censurés. L'hypothèse de log-normalité des résidus, déjà évoquée, est une source supplémentaire de mauvaise spécification. Si les queues étaient en réalité plus épaisses, la censure conduirait à l'obtention in fine de patrimoines plutôt trop faibles pour les plus patrimoines les plus importants.
La censure a également un effet sur la précision des estimateurs. La surreprésentation des ménages « aisés » est en effet destinée à améliorer cette précision. Le principe en est assez simple. En l'absence d'une telle surreprésentation, la probabilité de sélectionner un ménage aisé dans l'échantillon est très faible compte tenu de la forte concentration du patrimoine. L'échantillon contiendrait donc peu de ces ménages. Or la distribution du patrimoine reste très dispersée même au delà du dernier centile. Si l'opération était répétée, le sous échantillon de ménages aisés sélectionné pourrait avoir un patrimoine moyen par exemple très différent. La surreprésentation diminue cet aléa lié au sondage et améliore donc la précision des estimateurs de sondage.
Les bonnes propriétés d'un tel échantillonnage sont cependant en partie perdues du fait de la censure sur la variable récapitulative. Il est impossible de distinguer un ménage possédant 451 000 € d'un autre ménage multimillionnaire. La perte d'efficacité liée à la censure dépend en fait de la qualité de la surreprésentation. Elle peut être importante si les variables de surreprésentation sont bien corrélées au

[5] La loi des patrimoines ne doit toutefois pas être confondue avec la loi des résidus dans notre modèle. La première loi étant obtenue à partir de la seconde par marginalisation, dépend donc de la distribution des covariables.
[6] Selon la typologie construite par Nicole Tabard sur les données du recensement de 1999.

niveau de patrimoine et permettent effectivement d'augmenter la probabilité de sélectionner dans l'échantillon des ménages possédant bien plus de 450 000 €. Dans le cas de l'enquête Patrimoine, cet effet n'est donc peut-être pas si important. Il pourrait le devenir si la surreprésentation se faisait sur la base d'informations bien mieux corrélées au niveau de patrimoine, par exemple des informations fiscales comme celles issues des déclarations à l'impôt sur la fortune[7] ou de déclarations de revenus (notamment revenus du patrimoine)[8].

Des composantes au patrimoine total : une information plus riche mais plus complexe à utiliser

La censure limite de manière assez forte l'information sur le patrimoine total contenu dans la variable récapitulative pour les ménages se plaçant dans cette tranche. Or l'enquête procède à un recensement très complet et très détaillé des différentes composantes patrimoniales (immobilier, financier, professionnel). Cette décomposition doit permettre d'obtenir une information plus précise sur les ménages aisés. Même en raisonnant sur des montants observés seulement partiellement sous forme de tranches ou de fourchettes, il est possible de calculer des bornes inférieures pour le patrimoine total (construit comme la somme de ses composantes) qui dépassent le seuil de 450 000 €. Au delà du surcroît d'information provenant du niveau de détail, les caractéristiques des produits sont autant d'informations qu'il est possible d'ajouter aux caractéristiques ménages. Simuler le patrimoine à un niveau plus fin doit donc permettre, sous réserve de parcimonie, d'améliorer la couverture des intervalles de confiance des indices d'inégalités et éventuellement de corriger un biais.

L'information disponible dans l'enquête Patrimoine

L'enquête Patrimoine collecte une information très détaillée sur l'ensemble des éléments de patrimoine des ménages français. Pour le seul patrimoine financier, plus de 30 types de produits différents sont recensés : des livrets d'épargne règlementée aux valeurs mobilières en passant par les livrets soumis à l'impôt (livrets B, livrets Orange), les produits d'assurance-vie et d'épargne-retraite, ceux d'épargne-logement, ou encore d'épargne salariale. En plus de fournir des montants pour chacune de ces composantes. Le questionnaire recense également le patrimoine immobilier de jouissance (résidence principale, secondaire…) et de rapport (logements mis en location), ainsi que le patrimoine professionnel, exploité ou non par le ménage. Deux questions récapitulatives peuvent aussi être utilisées : une tranche est collectée pour la somme des composantes du patrimoine financier, la tranche pour le patrimoine total discutée dans la partie précédente. L'enquête de 2004 a par ailleurs été appariée avec des informations fiscales notamment sur l'imposabilité à l'ISF (mais pas sur le patrimoine imposable). La manipulation des inégalités liées aux contraintes sur la somme des parties ou de sous-parties (patrimoine financier) et l'imposabilité à l'ISF ont souvent permit d'affiner les intervalles contenant les diverses composantes.

Un modèle « univarié » utilisant des informations détaillées

Une première stratégie d'utilisation des informations détaillées consiste à procéder à des simulations comme dans la partie précédente dans un cadre univarié mais au niveau de chacune des composantes[9].
Notons que les informations sur le patrimoine total et l'imposabilité à l'ISF ont également permit de construire des intervalles contenant la composante non recensée : mélange de biens durables et de luxe (bijoux, œuvres d'art, collections…). Deux groupes ont été constitués afin de séparer le mieux possible les biens durables de ceux de luxe.
Une hypothèse simplificatrice, compte tenu du nombre élevé de composantes, est alors de supposer l'indépendance des résidus des différents modèles. Néanmoins, nous avons procédé à des simulations de l'ensemble des composantes de patrimoine, conditionnelles à l'observation des covariables tronquées. Le domaine de troncature est lui complexe si nous voulons tenir compte des informations mobilisant plusieurs variables. Le domaine de troncature est une partie de l'hyper-rectangle défini par les intervalles

[7] C'est la méthodologie utilisée pour l'enquête sur le patrimoine réalisée en Espagne par la Banca d'Espana (Encuesta Financiera de las Familias).
[8] C'est la méthodologie utilisée pour l'enquête sur le patrimoine réalisée aux Etats-Unis par la FED (Survey on Consumer Finance).
[9] Voir Annexe 1 pour plus de détail sur les modèles retenus pour les différentes composantes.

pour les composantes éventuellement affinés. Les trois contraintes supplémentaires permettent en effet de définir des bandes. Le domaine de troncature correspond donc à l'intersection de ces trois bandes et de l'hyper-rectangle. Nous avons procédé par acceptation/rejet (Robert et Casella (2004)). Les composantes ont dans un premier temps été simulées dans des lois tronquées indépendantes[10] (ce qui correspond à la troncation dans l'hyper-rectangle), puis nous avons accepté ces valeurs si elles satisfaisaient les contraintes imposées par la somme des parties, sinon nous avons recommencé à partir de nouveaux tirages de vecteurs indépendants (Annexe 2).
L'utilisation des composantes détaillées a très clairement l'avantage de pouvoir mieux appréhender des patrimoines qui dépassent la limite supérieure d'observation de 450 000 € pour la seule variable récapitulative, réduisant ainsi l'effet de cette censure sur l'éventuel biais ainsi que sur la précision.
Il existe toutefois aussi des limites à cette approche. A chaque étape de l'algorithme d'acceptation/rejet, nous faisons comme ci les composantes étaient indépendantes (conditionnellement aux observables x_k). Tenir compte des informations agrégées (récapitulatif du patrimoine financier, du patrimoine brut total, de l'imposabilité à l'ISF) dans l'algorithme de simulation permet certes d'avoir *in fine* un modèle où les montants de patrimoine alloués entre les différentes composantes sont dépendantes. Cependant, il est critiquable que non-conditionnellement à ces informations agrégées, les composantes soient indépendantes les unes des autres. En effet, les caractéristiques observables des ménages ou des produits n'expliquent qu'une part limitée de la variance des patrimoines. Des variables comme l'âge, le revenu et la catégorie sociale, qui sont les plus explicatives, peinent à expliquer plus de 30% de la variance du patrimoine total brut (Cordier, Houdré et Rougerie (2006)). Les préférences individuelles comme l'aversion pour le risque ou les préférences temporelles, auxquelles les développements récents de la théorie du cycle de vie confèrent un rôle important dans l'explication des comportements d'accumulation patrimoniale, sont difficiles à observer et à bien mesurer. En 1998, un sous-échantillon de répondants à l'enquête Patrimoine avait participé une enquête complémentaire[11] spécialement dédiée à la mesure directe de ces préférences. Arrondel, Masson et Verger (2005b) montrent à l'aide de ces sous-échantillons que la prise en compte de ces paramètres améliore l'explication de la variance des patrimoines (l'augmentation du R^2 après introduction de ces variables explicatives s'élève à 0.12) mais près de 45% de la variance reste malgré tout inexpliquée. L'existence de cette forte hétérogénéité inobservée (des capacités cognitives, à l'esprit d'entreprise par exemple) nous semble être une source de corrélation potentielle entre les composantes patrimoniales qui ne peut être prise en compte qu'en choisissant d'imputer des montants continus dans un cadre résolument multivarié.

Inférence à partir d'un modèle véritablement multivarié

Notre approche de l'inférence des inégalités est fondée sur un modèle hiérarchique comprenant un modèle pour les grandeurs d'intérêt en population totale et le DGP (Encadré 1 et 2). En raison du choix de l'échantillonnage de Gibbs pour la simulation, il est plus aisé d'adopter le point de vue de la statistique Bayésienne que le point de vue fréquentiste de la première partie. Par ailleurs cela permet de fournir des intervalles de confiance tenant compte de l'incertitude sur les paramètres. Une approche Bayésienne ajoute un « étage » à notre modèle hiérarchique, la spécification de la loi a priori des paramètres (Encadré 3).
mettant l'inférence
Nous procédons à une modélisation jointe relâchant l'hypothèse d'indépendance entre les montants des composantes détenues par les ménages. Nous devons rester relativement parcimonieux dans le nombre de composantes et de variables explicatives en l'absence de restrictions sur la matrice de variance-covariance des résidus des composantes et contrairement à l'approche de la partie précédente, les composantes sont agrégées. Nous distinguons

1. Le patrimoine financier (FIN)
2. Le montant de la part possédée de la résidence principale (RP)
3. Le montant de la part possédée des autres biens immobiliers : autres logements et parkings (ALG)
4. Le montant de la part possédée de patrimoine professionnel (PROF)

[10] La simulation de telles lois univariées tronquées est facile en utilisant la fonction quantile de la loi normale et la fonction de répartition disponible dans les logiciels statistiques ou l'acceptation-rejet avec une loi instrumentale bien choisie, cf. Robert (1995).
[11] L'enquête Insee-Delta, voir Arrondel, Masson et Verger (2005a).

5. Le reste : biens durables, œuvres d'art, bijoux, collections privées (RESTE). Cette dernière n'est observée qu'indirectement à l'aide de la variable récapitulative totale (cf. Annexe 1).

Nous faisons l'hypothèse simplificatrice que tous les ménages possèdent du patrimoine financier (qui comprend les encours sur les compte-chèques), ainsi que du patrimoine sous forme de biens durables, bijoux ou objets d'art. Il existe alors seulement $2^3 = 8$ types de portefeuilles différents.

La résidence principale est la seule composante qui, par rapport aux données d'enquête, n'est pas agrégée. Il a été possible d'utiliser des covariables produit (surface, surface au carré…). Ainsi, alors que pour les autres composantes le modèle porte sur la part possédée par le ménage, en ce qui concerne la résidence principale, le modèle porte sur le montant lui-même. Nous donnons les variables que nous avons retenu, hors effets fixes groupes, pour les modèles, à un niveau plus ou moins regroupé.

Tableau 2 : Variables explicatives retenues pour chaque composante

Covariables \ Composantes	**FIN**	**RP**	**ALG**	**PROF**	**RESTE**
Cycle de vie					
seul et sans enfant		X	X	X	X
âge et âge au carré		X	X	X	X
position dans le cycle de vie	X				
ressources culturelles et sociales					
niveau social	X	X	X	X	X
niveau d'études	X	X	X	X	X
Revenus					
niveau de salaire	X	X	X	X	
perception d'aides sociales	X				
perception d'une rente	X	X		X	
perception de revenus autres que revenus d'activité ou de remplacement	X		X	X	
Zone géographique	X	X	X		X
Histoire du patrimoine					
existence d'une donation reçue	X	X	X		X
existence d'une donation versée	X				
augmentation ou baisse récente du patrimoine	X	X		X	X
composition patrimoine des parents	X		X	X	
Résidence principale					
surface et surface au carré		X			
Patrimoine professionnel					
existence de patrimoine professionnel				X	
possession d'une entreprise				X	

Encadré 3

Le modèle Bayésien multivarié hiérarchique[12]

I - Loi des grandeurs d'intérêt (I)

Les grandeurs d'intérêt sont modélisées par un système de relations du type (cf. Encadré 1) :

$$G = \hat{G}(pt_1, \dots, pt_m) + \sqrt{\hat{V}(\hat{G})(pt_1, \dots, pt_m)}\epsilon, \text{où } \epsilon \text{ suit une loi normale centrée réduite,}$$

le patrimoine total pt_k du ménage numéroté k est donné par $pt_k = \sum_{l=1}^{5} P_k^l$, P_k^l est la $l^{ème}$ composante de patrimoine du ménage: $P_k^1 = FIN_k, P_k^2 = RP_k, P_k^3 = ALG_k, P_k^4 = PROF_k$ et $P_k^5 = RESTE_k$.

Nous définissons $D_k^l = 1\{P_k^l > 0\}$, la variable valant 1 si le $k^{ème}$ ménage a du patrimoine en la $l^{ème}$ composante et 0 sinon. Le vecteur $\boldsymbol{D_k}$[13] de taille 5 est le vecteur de décomposition du patrimoine.

[12] Il ne s'agit pas d'une approche Bayésienne hiérarchique pour laquelle nous poserions une loi sur les hyper-paramètres de la loi a priori. Ici la loi a priori étant non informative, il n'y a pas d'hyper-paramètres.

[13] Nous ne modélisons pas la loi des $\boldsymbol{D_k}$, ils sont observés et nous travaillerons conditionnellement à l'observation de leurs valeurs. Nous avons également spécifié un modèle Tobit multivarié. Il donne des résultats peu crédibles. Il est de toute façon restrictif qu'une variable rende compte conjointement de la détention et du montant. Tout comme il n'est pas utile de spécifier un modèle en

II - Le processus de génération des données (DGP)

Nous modélisons le vecteur des patrimoines détenus par chaque ménage de l'échantillon (nous faisons conjointement l'hypothèse *(H)*, voir plus bas).

Soit P la fonction qui à $\boldsymbol{D_k}$ associe un entier dans $\{1, \dots, 8\}$. Nous spécifions alors 8 modèles différents conditionnels à la valeur prise par $P(\boldsymbol{D_k})$. $P(\boldsymbol{D_k}) = 1$ correspond au portefeuille tel que le ménage possède les 5 composantes de patrimoine (chaque composante de $\boldsymbol{D_k}$ vaut 1).

Sachant $P(\boldsymbol{D_k}) = i$, nous avons le système de $d_i = \sum_{l=1}^{5} D_k^l$ équations simultanées tel que:

$$\log\left(P_k^l\right) = x_{l,k}\beta_l^i + u_l^i, lorsque\ D_k^l = 1$$

où le vecteur $\boldsymbol{u^i}$ des résidus suit la loi d'un vecteur Gaussien centré de matrice de variance-covariance Σ^i de taille $d_i * d_i$.

Hypothèse (ignorabilité ou sélection exogène)

(H) Conditionnellement à l'observation des covariables des modèles, ϵ est independent de $(pt_1, \dots, pt_m)$

III – L'a priori[14] (P)

Par souci de parcimonie, nous posons (restriction sur les paramètres) que les vecteurs β_l^i ont les mêmes composantes pour tous les indices i de groupes tels que $D_\cdot^l = 1$ hormis la première coordonnée qui est spécifique aux groupes, effet fixe groupe[15].

Si nous notons par dim_l la dimension des vecteurs β_l^i (égaux pour tout i) et de même pour les 5 autres composantes le vecteur des paramètres θ est de dimension

$$\sum_{l=1}^{5}(dim_l - 1) + 8 * 5 + \frac{1}{2}\sum_{k=2}^{5} k(k+1).$$

Nous prenons pour la densité $\pi(\theta)$ de θ la fonction proportionnelle à

$$\prod_{i=1}^{8} \det\left(\Sigma^i\right)^{-(p_i+1)/2}.$$

où p_i correspond au nombre de lignes de Σ^i.

Remarque : Cette loi est conjuguée[16] au DGP. Elle est impropre, c'est-à-dire qu'il ne s'agit pas d'une probabilité car la fonction n'est pas intégrable et donc son intégrale ne vaut pas 1. Les coefficients des paramètres des droites de régression sont pris tous a priori indépendants et la loi a priori du vecteur est proportionnelle à la mesure de Lebesgue dans $\mathbb{R}^d$ où $d = \sum_{l=1}^{5}(dim_l - 1) + 8 * 5$. L'a priori est donc tout à fait objectif car nous ne faisons aucune hypothèse sur coefficients. Ils sont par ailleurs a priori indépendants des matrices de variance-covariance, elles aussi sont prises a priori mutuellement indépendantes. Il est classique qu'il n'est pas possible de prendre une loi diffuse pour les matrices Σ. Bien que $\pi(\theta)$ ne soit pas associé à une probabilité, $\pi(\theta|P_k^l, x_{l,k}, k = 1, \dots, m, l)d\theta$ est bien une probabilité. Dans le cas où il n'y aurait eu qu'un seul groupe (disons le groupe 1), cette loi a priori est très usuelle dans les modèles linéaires Gaussiens multivariés, cf. Little et Rubin (2002) et Schafer (2001). Elle est parfois dite non-informative ou de loi de Jeffreys. Il s'agit d'une limite de lois normale / inverse-Wishart, lorsque $\tau \to 0, m \to -1, \Lambda^{-1} \to 0$, définie par :

1. $\beta = \left(\beta_1^{1'}, \dots, \beta_5^{1'}\right)'$ suit une loi normale $\mathcal{N}\left(\beta_0, \tau^{-1}\Sigma^1\right)$,

---

population générale (il suffit de modéliser les sélectionnés sous l'hypothèse *(H)*), modéliser la détention est inutile quant à l'inférence des grandeurs d'intérêt. Néanmoins nous tenons compte de la détention via les effets fixes.

[14] Le vecteur des paramètres n'est pas supposé avoir une vraie valeur. En l'absence de données nous avons une connaissance « vague » sur le paramètre, voire aucune, on rend compte de l'incertitude en spécifiant une loi de probabilité pour la position du paramètre : la loi a priori. Lorsque des données arrivent, un modèle de type DGP étant spécifié, il est possible d'avoir une connaissance plus précise sur le paramètre : sa loi a posteriori, loi conditionnelle à l'observation des données. Cette loi se met à jour à l'arrivée de chaque donnée complémentaire, la loi a posteriori devenant loi a priori. Cette loi se concentre à mesure que la taille de l'échantillon grandit. Un avantage est donc que la statistique Bayésienne s'accommode d'échantillons finis et permet de tenir d'un possible a priori, néanmoins l'a priori est aussi vu par beaucoup comme un désavantage car sa spécification n'est pas aisée. L'imputation Bayésienne multiple (Little et Rubin (2002)) repose sur ce point de vue Bayésien et essentiellement motivé par le fait qu'il permet de tenir compte de l'incertitude sur les paramètres.

[15] Nous avons pour cela ajouté comme covariables des indicatrices d'appartenance aux groupes.

[16] Dans ce cas la loi a posteriori $\pi(\theta|Y = y)d\theta$ est dans la même famille paramétrée de lois que $\pi(\theta)d\theta$.

2. La loi de Σ^1 sachant β est inverse-Wishart[17] $\mathcal{W}^{-1}(m, \Lambda)$

La censure

A partir des intervalles pour les parties des composantes constituant nos 5 composantes agrégées et des informations agrégées (récapitulatif du patrimoine financier, du patrimoine brut total, de l'imposabilité à l'ISF), qui permettent d'affiner ces intervalles (cf. Annexe 1), nous savons que les vecteurs des composantes de patrimoine détenu de dimension d_i sont dans des hyper-rectangles.
La question récapitulative par exemple nous donne par exemple que ces vecteurs sont dans l'intersection de l' hyper-rectangle et d'une bande définie par

$$pt_{k,min} \leq \sum_{l=1}^{5} R_k^l D_k^l P_k^l \leq pt_{k,max}$$

où R_k^l est la part possédée par le ménage, nous avons choisi systématiquement de modéliser la part possédée par le ménage sauf pour la résidence principale, tous ces coefficients valent 1 sauf pour la résidence principale.
L'imposabilité à l'ISF du ménage permet de restreindre à nouveau le domaine de censure. La base du patrimoine utilisée pour le calcul de l'ISF est moins large que le concept retenu dans l'enquête : le patrimoine professionnel n'est pas entièrement pris en compte (par exemple si on ne possède qu'une part trop faible d'une entreprise, il n'est pas possible de déduire ce montant du calcul du patrimoine imposable), la résidence principale bénéficie d'un abattement de 20%, les objets d'art ne sont pas non plus imposés. La décomposition du patrimoine total retenue distingue la résidence principale des autres actifs immobiliers ou fonciers, ainsi que le patrimoine professionnel mais sans faire la distinction entre le patrimoine professionnel exploité pouvant faire l'objet d'une défiscalisation, du reste du patrimoine professionnel, exploité ou non. Il a été possible de tenir compte du patrimoine professionnel exploité et en particulier des entreprises car le peu de personnes en possédant avaient rempli un montant en clair. Nous sommes capables de produire des bornes supérieures ou inférieures liées au seuil d'imposition à l'ISF qui en 2003 s'élevait à 720 000 €. Nous avons les deux situations suivantes :

1. Lorsqu'un ménage est imposé à l'ISF, son patrimoine imposable est supérieur à 720 000 €. La borne supérieure suivante du patrimoine imposable
$$FIN_k + 0.8 * RP_k + ALG_k + MIN\left(PROF_k, NDED_{max,k}\right) + RESTE_k - PASSIF_k$$
doit donc être supérieure à 720 000 €. $NDED_{max,k}$ correspond à une borne supérieure de la valeur maximale de l'ensemble du patrimoine professionnel non déductible construite à partir des informations détaillées et $PASSIF_k$ au passif qui est déductible et nous supposons qu'il est constamment déduit.
2. Lorsqu'un ménage n'est pas imposé à l'ISF, son patrimoine imposable est inférieur à 720 000 €. La borne inférieure suivante du patrimoine imposable
$$FIN_k + 0.8 * RP_k + ALG_k + NDED_{min,k} - PASSIF_k$$
doit donc être inférieure à 720 000 €. $NDED_{min,k}$ est une borne inférieure du patrimoine professionnel non déductible construite à partir des informations détaillées.

Enfin, des informations externes sur la distribution du patrimoine ont également servi à limiter le niveau de patrimoine qui pouvait être simulé. Ce choix, discutable, résulte de ce que l'enquête n'étant réalisée qu'une seule fois, il est possible qu'ou bien nous n'ayons aucun patrimoine vraiment élevé ou que nous en ayons au moins un « trop élevé ». En effet, à chaque ménage correspond un poids de sondage, si le patrimoine d'un milliardaire est recueilli, il se verra en moyenne affecté un poids de l'ordre du millier et en quelque sorte représentera 1 000 ménages faisant exploser les inégalités. Mais en général, c'est plutôt l'inverse qui se produit, ceux-ci sont tout de même assez rares. Voila pourquoi il a été décidé de surreprésenter certaines catégories dans l'espoir d'observer des patrimoines élevés (avec du fait de la surreprésentation un poids plus faible, ce qui a donc toute les propriétés souhaitables en terme de gain de précision sondage si on ne procédait pas à une collecte par intervalles). Même s'il n'est pas possible

[17] La loi de Wishart est l'extension aux matrices symétriques réelles définies positives de la loi du χ^2. Elle a la même distribution que la loi des matrices $X'X$ ou X est une matrice à m lignes et p colonnes dont les lignes sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \Lambda)$. On peut trouver un algorithme efficace de simulation dans McCulloch et Rossi (1994).

en général de distinguer un ménage possédant plus de 450 000 € d'un milliardaire, nous sommes dans un cas somme toute favorable. Un des ménages enquêtés a parfaitement renseigné la valeur de son patrimoine professionnel et nous disposons donc d'un patrimoine relativement élevé, autour de 25 000 000 € et il ne semble pas que ne disposions de patrimoines très élevés. En moyenne toutefois, les ménages enquêtés représentent 2000 ménages français. Les poids de sondage varient entre 450 et 12 000. Nous avons donc limité la possibilité que le patrimoine simulé dépasse une valeur au-delà de laquelle le principe de représentativité soit mis en défaut. Sur la base de Cordier et al. (2006) et d'informations publiques sur les plus grosses fortunes professionnelles françaises[18], nous avons introduit des plafonds, relativement généreux sur le patrimoine total. Le patrimoine de notre ménage, apparemment le plus fortuné, a été plafonné à 50 millions d'euros et celui des autres ménages à 10 millions d'euros. On peut par contre s'attendre à ce que l'introduction de telles restrictions conduise à sous-estimer légèrement la couverture des intervalles de confiance. Heeringa, Little et Raghunatan (2002) constatent avec un modèle analogue et sur les données de l'enquête HRS que l'introduction de telles restrictions conduit à minorer des indicateurs comme la moyenne, l'indice de Gini ou d'autres indices de concentration. En revanche, l'impact sur des indicateurs plus robustes tels que les quantiles est minime.

L'échantillonnage de Gibbs : simuler dans la loi jointe des composantes tronquée en restant dans un cadre univarié et facilité de l'inférence avec une approche Bayésienne.

L'échantillonnage de Gibbs[19] (Arnold (1993) ou Robert (1995) et Encadré 4) est un algorithme de simulation particulièrement efficace pour simuler des vecteurs aléatoires. Son utilisation dans le cas de vecteurs Gaussiens tronqués est par exemple proposé dans Robert (1995). De tels vecteurs aléatoires sont effectivement difficiles à simuler de façon efficace lorsque la dimension du vecteur et du domaine de troncature sont élevées[20].

Encadré 4

Principe de l'échantillonnage de Gibbs

Nous rappelons ici brièvement le principe de la méthode sur un exemple ou le vecteur et découpé en 2 sous-vecteurs, le cas d'un découpage en un nombre plus élevé de composantes est analogue. Supposons que nous souhaitons produire un tirage dans la loi d'un vecteur aléatoire $X = (X^{1\prime}, X^{2\prime})'$ et qu'il soit aisé de simuler dans les lois conditionnelles notées $\mathcal{L}(X^1|X^2 = x^2)$ et $\mathcal{L}(X^2|X^1 = x^1)$, dans ce cas, partant d'une condition initiale $x_0 = (x_0^{1\prime}, x_0^{2\prime})'$, on construit une suite de vecteurs $x_n = (x_n^{1\prime}, x_n^{2\prime})'$ telle que connaissant x_{n-1} on tire successivement :

(i) x_n^1 dans $\mathcal{L}(X^1|X^2 = x_{n-1}^2)$,
(ii) x_n^2 dans $\mathcal{L}(X^2|X^1 = x_n^1)$.

La simulation d'un vecteur est décomposée en des simulations de vecteurs de dimension inférieure, voir de variables aléatoires. Nous construisons une trajectoire de chaîne de Markov, à temps discret et espace d'états continu, cf. Meyne et Tweedie (1993) de probabilité invariante μ : loi $\mathcal{L}(X)$ du vecteur X. Les deux propriétés suivantes seront satisfaites et reposent sur le caractère Markovien :

l'ergodicité : quelque soit la loi de X_0,les lois des marginales $\mathcal{L}(X_n)$ convergent vers μ ;

le théorème ergodique : soit f mesurable telle que $\mathbb{E}_\mu[|f(X)|] < \infty$ alors[21] :

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T} f(X_n) = \mathbb{E}_\mu[f(X)] \text{ p. s. pour toute donnée initiale .}$$

18 Le magasine Challenges publie chaque année un classement des 500 plus grosses fortunes professionnelles de France.

19 Il s'agit d'une méthode de Monte Carlo par chaine de Markov (MCMC), voir Robert et Casella (2004).

20 Dans le cas du maximum de vraisemblance simulé, la simulation n'est pas effectuée dans la loi du vecteur Gaussien tronqué pour des questions de régularité et de variance. Il est plus efficace de procéder par échantillonnage préférentiel, qui permet de simuler dans d'autres lois, éventuellement plus faciles. Une loi souvent utilisée est celle des Gaussiennes tronquées itérées à l'aide de l'algorithme GHK (Geweke, Hajivassiliou et Keane).

21 Il s'agit d'une loi des grands nombres dans un cas de variables aléatoires dépendantes.

Les théorèmes limites dans le cas particulier de l'échantillonnage de Gibbs sont donnés dans Tierney (2004) et Roberts et Smith (1994). Remarquons, au sujet de l'ergodicité, que comme dans Roberts et Polson (1994), en minorant le noyau de transition et en introduisant des bornes supérieures (même lâches) pour les montants non bornés, la convergence est exponentielle.

Si l'on décide d'utiliser l'échantillonnage de Gibbs pour la simulation, l'extension à une inférence basée sur l'approche Bayésienne (Robert (2007)) est relativement naturelle et plus facile qu'une approche fréquentiste. L'estimation d'un modèle multivarié en présence de censure[22] est un problème intensif du point de vue du calcul[23]. Il existe certes de nombreuses méthodes comme le maximum de vraisemblance simulé, les moments ou scores simulés ou des variantes de l'algorithme MC-EM. En outre, les fonctionnelles à optimiser présentent en général de multiples extrema locaux et il est souvent utile d'employer un algorithme d'optimisation stochastique, ou au moins d'initier la méthode déterministe avec plusieurs points différents. Ces paramètres difficiles à obtenir ne sont d'ailleurs pas les grandeurs cibles ici et il faut ensuite simuler dans les lois multivariées tronquées afin in fine de produire des intervalles de confiance des indicateurs d'inégalités pour la distribution du patrimoine total de l'ensemble des français, tenant si possible compte de l'incertitude sur les paramètres liés à la taille finie de l'échantillon.

L'échantillonnage de Gibbs est utilisé sur le vecteur de très grande dimension $V = \left((\boldsymbol{P_1'}, \dots, \boldsymbol{P_m'}), \theta', \epsilon\right)'$ où $\boldsymbol{P_i}$ sont les vecteurs de composantes de patrimoine détenues par le ménage i, ils sont indépendants entre eux, et ϵ est le terme d'erreur indépendant de θ et $(\boldsymbol{P_1'}, \dots, \boldsymbol{P_m'})'$ apparaissant dans la formule (I) et les étapes de type (i) et (ii) sont aisées. Nous découpons par ailleurs chaque $\boldsymbol{P_k}$ en sous-vecteurs de taille 1. Les lois des $\boldsymbol{P_k}$ sont tronquées et conditionnelles à l'observation de covariables. Dans ce cas, nous nous sommes soustraits des calculs itérés de nombreuses (m) intégrales multiples et des procédures d'optimisation complexes. En contrepartie, le vecteur à simuler par échantillonnage de Gibbs est de taille plus importante. L'inférence sur les indicateurs d'inégalité se fait via la formule (I). L'inférence par intervalles se fait comme dans l'Encadré 2 en substituant la loi des grands nombres par le théorème ergodique en remplaçant f par G dans la formule. Le risque *(R)* est conditionnel à l'observation des covariables et des censures. Nous devons donc effectivement simuler dans des lois tronquées.

Du point de vue de la mise en œuvre pratique nous procédons comme suit. Nous nous donnons une donnée initiale pour le sous-vecteur de l'ensemble des composantes de patrimoine des m ménages satisfaisant l'ensemble des contraintes. En ce qui concerne une condition initiale θ_0 pour θ, il n'est pas nécessaire d'initier les paramètres des droites de régression car leur loi conditionnelle aux données et covariances ne dépend pas des données initiales que via les composantes de patrimoine des m ménages et les matrices $\Sigma_0^1, \dots \Sigma_0^8$ que nous initions comme des matrices diagonales où les termes diagonaux sont les variances des résidus dans des modèles pour les marginales. Décrivons la mise à jour du vecteur X à l'étape 1. Nous procédons de même aux étapes suivantes :

Mise à jour du vecteur $\boldsymbol{\beta}$: désignant ici l'ensemble des coefficients apparaissant dans les moyennes, en fonction des matrices $\Sigma_0^1, \dots \Sigma_0^8$: nous calculons

$$\Sigma_\beta = \left(\sum_{k=1}^{m} \boldsymbol{X_k'} \left(\Sigma_0^{P(k)}\right)^{-1} \boldsymbol{X_k} \right)^{-1}$$

où $\boldsymbol{X_k}$ est la matrice diagonale par blocs où figurent sur la diagonale les lignes de covariables pour chaque composante de patrimoine détenue pour le ménage k, et

$$\hat{\beta} = \Sigma_\beta \left(\sum_{k=1}^{m} \boldsymbol{X_k'} \left(\Sigma_0^{P(k)}\right)^{-1} \boldsymbol{P_0^k} \right)$$

où $\boldsymbol{P_0^k}$ correspond aux données initiales des composantes de patrimoine détenue pour le ménage k, enfin nous simulons β_1 dans la loi $\mathcal{N}\left(\hat{\beta}, \Sigma_\beta\right)$.

Mise à jour des matrices de variance-covariance : il convient de construire les matrices $\Lambda_1^1, \dots, \Lambda_1^8$ définissant les matrices $\Sigma_1^1, \dots, \Sigma_1^8$, si on note Π_i la projection qui permet de passer de β_1 au sous-vecteur pertinent dans le cas d'un portefeuille de type i, Λ_1^i est donné par

[22] Le domaine est ici en outre non rectangulaire.

[23] Pour un aperçu dans le cadre du Probit multivarié on peut consulter Train (2003).

$$\Lambda_1^i = \sum_{k:P(k)=i} \left(\boldsymbol{P_0^k} - \boldsymbol{X_k}\Pi_i\beta_1\right)\left(\boldsymbol{P_0^k} - \boldsymbol{X_k}\Pi_i\beta_1\right)'.$$

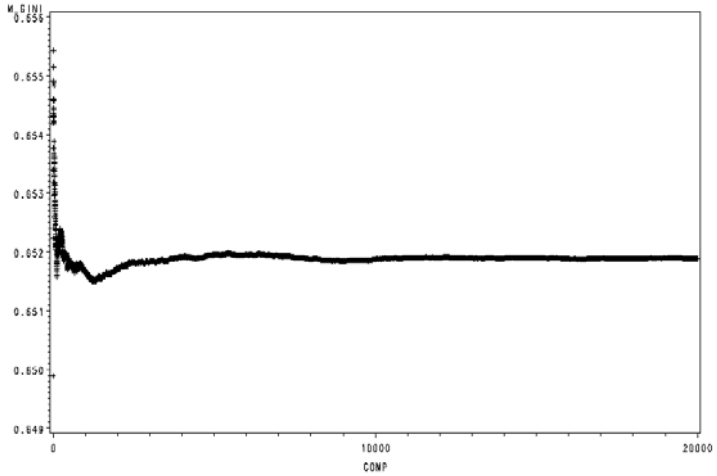
Simulation des composantes de patrimoine possédées pour l'ensemble des m ménages : Nous actualisons $(\boldsymbol{P_1'}, \ldots, \boldsymbol{P_m'})$ ménage après ménage et procédons pour chaque ménage composante par composante. Cette dernière décomposition en vecteurs de plus petite dimension permet d'utiliser des algorithmes très performants de simulation de variables tronquées en dimension 1, nous utilisons alors les formules usuelles de conditionnement dans un vecteur Gaussien. Nous effectuons des tirages conditionnels aux paramètres actualisés et aux valeurs actualisées des patrimoines au dessus dans la séquence ou leur valeur à l'étape 0 pour les patrimoines postérieurs dans la séquence (rappelons que les composantes sont dépendantes entre elles au sein de chaque ménage mais indépendantes d'un ménage à l'autre). Nous fabriquons également composante après composante la contrainte de troncature qu'elle doit satisfaire en mobilisant convenablement toutes les informations en intervalles et les valeurs précédemment simulées. Ces lois conditionnelles sont des lois normales tronquées[24]. Ceci est l'attrait majeur de l'échantillonnage de Gibbs pour simuler des vecteurs Gaussiens tronqués.
Ces formules des lois conditionnelles apparaissant dans l'échantillonnage de Gibbs s'obtiennent aisément en utilisant la formule de Bayes et la propriété que les lois normales et normale/inverse-Wishart sont conjuguées. Notons que si on procède de la sorte, en temps long (T grand) mais pas nécessairement trop long car la convergence est exponentielle, la loi marginale des paramètres est une bonne approximation de la loi a posteriori. De même, la loi marginale des composantes de patrimoines des m ménages est une bonne approximation de la loi dite prédictive a posteriori. Comme sous-produit de notre approche nous complétons notre fichier un nombre élevé de fois (imputations multiples) par simulations dans approximation de la loi prédictive a posteriori. Mais notre objet n'est pas ici l'imputation mais bien l'estimation des indicateurs d'inégalités de patrimoine total des ménages en France.

Intervalles de confiance et prévision

Le principe de l'inférence est une adaptation de la méthodologie de l'Encadré 2 au cas de données fournies par une trajectoire de chaine de Markov. Nous remplaçons la loi des grands nombres par le théorème ergodique. En théorie, il suffit de calculer des moyennes le long d'une trajectoire de V_n. Afin de vérifier que nous n'étions pas dans un cas pathologique, nous avons essayé quelques points de départ différents qui nous ont tous donné les mêmes résultats. Notons aussi qu'il est courant de ne pas utiliser la première fraction des simulations (*burn-in*). Toutes les représentations graphiques de la convergence des moyennes donnent une convergence très rapide, ce qui est conforme au résultat de convergence exponentielle vers l'équilibre, nous ne faisons figurer ici que le graphique pour l'indicateur de Gini.

[24] Cette propriété n'est pas vraie lorsque l'on considère des marginales de vecteurs Gaussiens tronqués ou que nous dévissons la loi jointe de chaque vecteur en lois conditionnelles itérées. Cette dernière approche est celle de l'algorithme GHK, qui au lieu de simuler itérativement dans les lois conditionnelles itérées, simule dans des gaussiennes tronquées mais le biais est corrigé via l'introduction dans l'approximation Monte-Carlo de poids d'échantillonnage préférentiel.

Graphique 3 : Convergence exponentielle de l'estimateur par prévision de l'indice de Gini



Source : Enquête Patrimoine 2004, calculs des auteurs.
Note : T=20 000.

Du fait de la rapidité de la convergence et comme nous effectuons un très grand nombre ($T = 20\,000$) de simulations, les calculs avec et sans *burn-in* sont quasiment indistinguables (Tableaux 3 et 4). Commençons par présenter les intervalles de confiance. Ils sont pris symétriques, nous aurions aussi pu les prendre de longueur minimale ou HPD (Higher Posterior Distribution).

Tableau 3 : Estimation Bayésienne par intervalle à 5% sans *burn-in* d'indicateurs d'inégalités

Grandeur d'intérêt	**Prévision**	**Intervalle de confiance**	
		Borne inférieure	**Borne supérieure**
Moyenne	204 995,03	192 879,51	217 653,77
Médiane	111 457,83	105 682,08	117 561,79
P99	1 584 469,11	1 358 490,05	1 823 925,98
P95	690 746,12	636 890,42	746 759,31
P90	434 455,79	416 410,51	452047,99
Q3	232 305,83	224 856,83	240 216,02
Q1	16 998,18	15 122,27	19 155,83
P10	3 958,92	2 873,70	5 073,87
P95D5	6,1968	5,7230	6,6805
P99D5	14,2164	12,1616	16,4277
Q3Q1	690 783,23	651 792,54	731 751,40
D9D1	109,9214	80,7967	140,5666
D9D5	3,8981	3,7086	4,0865
Gini	0,6519	0,6328	0,6717
Theil	0,9044	0,8138	1,0001
Atkinson ($\epsilon = 1.5$)	0,9063	0,8837	0,9253
Atkinson ($\epsilon = 2$)	0,9741	0,9549	0,9919

Source : Enquête Patrimoine 2004, calculs des auteurs.
Note : T=20 000.

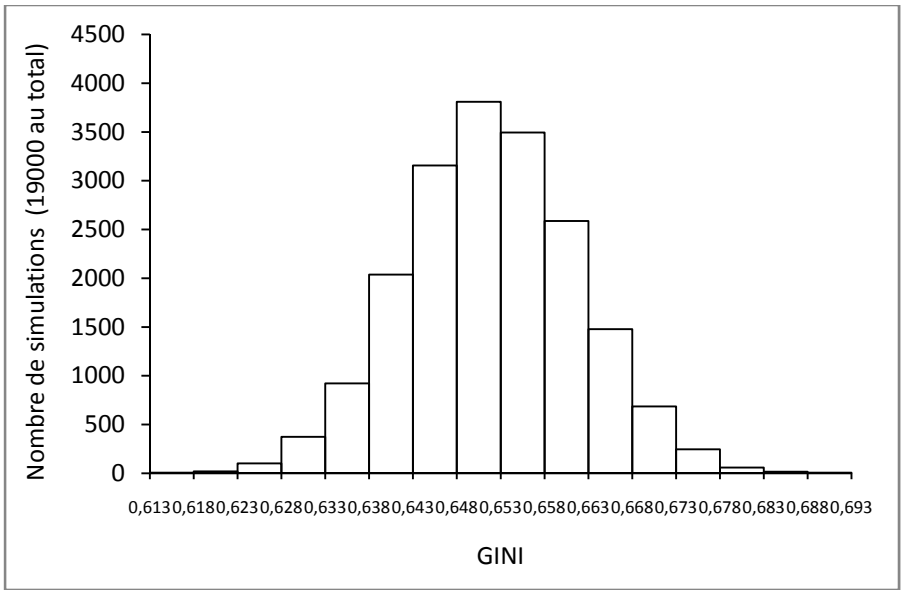
Tableau 4 : Estimation Bayésienne par intervalle à 5% avec *burn-in* d'indicateurs d'inégalités

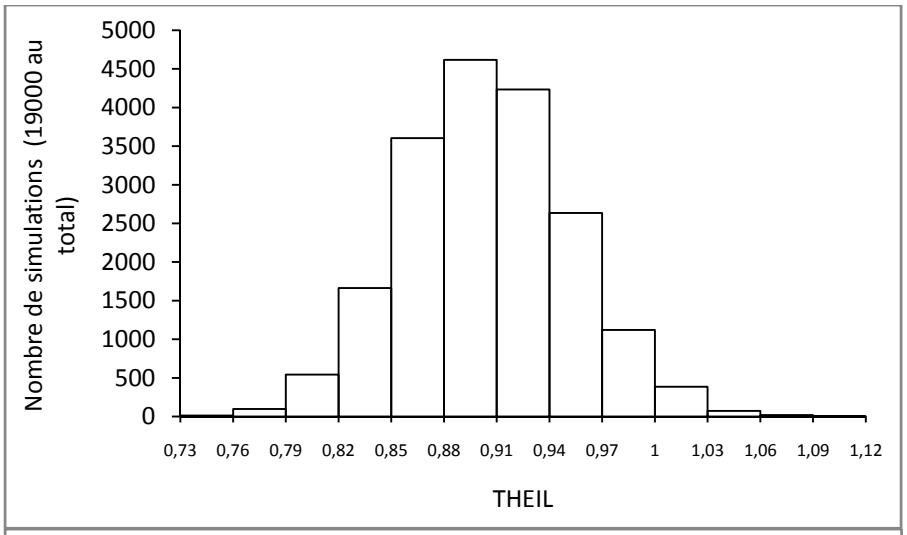
Statistique	Estimateur	Borne inférieure	Borne supérieure
Moyenne	205 003,98	192 879,50	217647,24
Médiane	111 459,26	105 672,32	117 563,45
P99	1 584 602,96	1 359 261,98	1 825 362,43
P95	690 793,96	636 924,50	746 759,31
P90	434 458,13	416 410,51	452 006,79
Q3	232 307,50	224 849,03	240 204,86
Q1	16 998,67	15 117,08	19 149,60
P10	3 959,07	2 870,83	5 070,55
P95/D5	6,1972	5,7232	6,6808
P99/D5	14,2175	12,1667	16,4388
Q3/Q1	13,6847	12,2838	15,1034
D9/D1	109,9332	80,7615	140,5827
D9/D5	3,8981	3,7081	4,0858
Gini	0,6519	0,6328	0,6717
Theil	0,9044	0,8138	1,0001
Atkinson ($\epsilon = 1.5$)	0,9063	0,8838	0,9253
Atkinson ($\epsilon = 2$)	0,9742	0,9549	0,9920

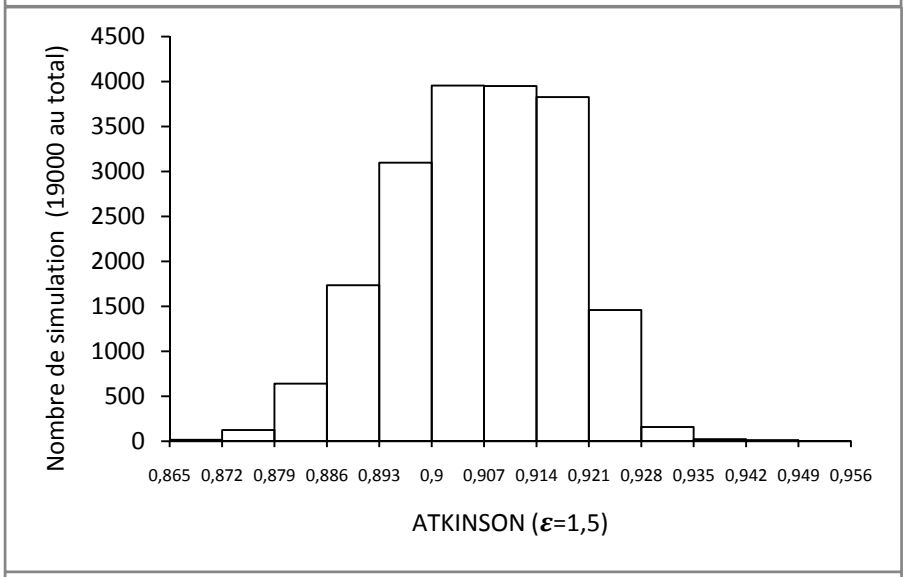
Source : Enquête Patrimoine 2004, calculs des auteurs.
Note : T=20 000, *burn-in* des 1 000 premières simulations.

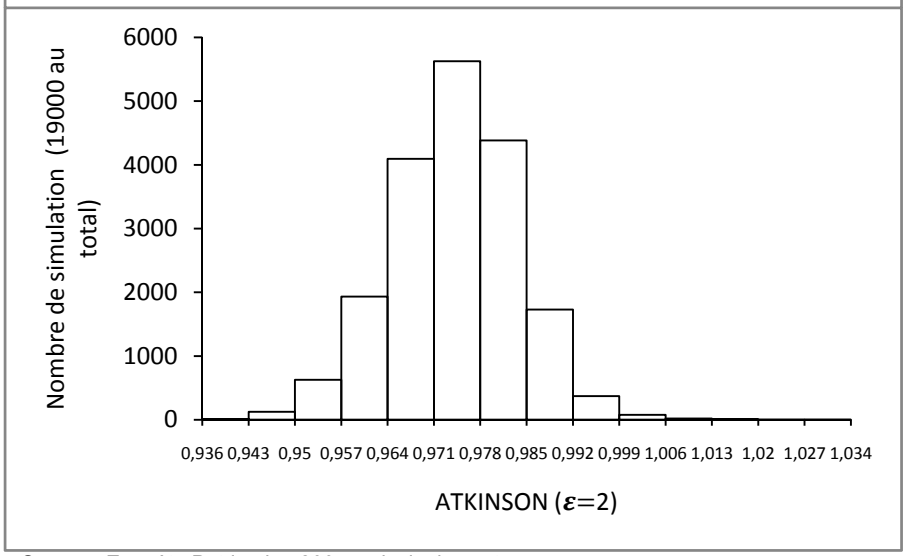
Nous obtenons aussi par exemple la distribution empirique (ici avec *burn-in*) suivante de la distribution a posteriori des grandeurs d'intérêt dans le cas des indicateurs de Gini, de Theil et d'Atkinson.

Graphique 4 : Distribution empirique approchant la loi a posteriori des grandeurs d'intérêt









Source : Enquête Patrimoine 2004, calculs des auteurs.
Note : T=19 000.

La comparaison des tableaux 1 et 4 permet d'appréhender comment se déforme la distribution des patrimoines suivant qu'elle est estimé à partir de données simulées avec la seule variable récapitulative, ou bien avec un modèle multivarié pour plusieurs composantes de patrimoine. Les indicateurs synthétiques de dispersion (rapports inter-quantiles) ou de concentration (Gini, Theil) montrent bien que la méthode multivariée conduit à des distributions plus dispersées. Le fait de n'utiliser que la variable récapitulative semble « tasser » la distribution à ses deux extrémités : les petits patrimoines (P10) sont supérieurs à ceux simulés dans l'approche multivariée, tandis que les gros patrimoines (P90, P95, P99) sont en revanche plus faibles. En outre la couverture des intervalles de confiance pour l'indice de Gini (par exemple) est à peu près la même. Or l'intervalle calculé avec la variable récapitulative ne tient pas compte de l'incertitude sur les paramètres.

Discussion

Notre modèle multivarié est voisin de celui utilisé dans Heeringa et al. (2002). Nous modélisons la loi jointe des composantes détenues parmi les 5. Nous posons alors un modèle linéaire Gaussien pour la loi jointe des logarithmes des variables et supposons que le vecteur des résidus possède une matrice de variance-covariance quelconque. Heeringa et al. (2002) considèrent une décomposition plus fine du patrimoine mais n'introduisent par contre quasiment aucune covariable. Ils autorisent autant de moyennes différentes que de compositions du portefeuille mais en revanche, les matrices de variance-covariance des résidus sont des blocs extraits d'une même matrice de dimension égale au nombre total de composantes possibles. La stabilité de la loi des paramètres après itération de la chaine ne nous semble pas clair et il nous semble par ailleurs que cette structure revient à penser que l'absence d'une composante est traitée comme si elle n'était pas observée mais existait ce qui ne nous parait pas justifié. Nous tenons également compte de la détention mais, par souci de parcimonie, nous imposons des contraintes sur les paramètres apparaissant dans les moyennes des logarithmes : seule la constante, effet fixe, est spécifique au type de portefeuille ; par contre, nous utilisons des matrices de variance-covariance spécifiques à chacun des 8 types de portefeuille. L'effet fixe est une première approche de l'hétérogénéité des groupes même si on peut regretter que l'on n'ait pas utilisé des vecteurs de coefficients par groupes. Il serait possible de considérer un plus grand nombre de composantes du patrimoine, en introduisant par exemple moins de covariables.
Une partie de la différence de résultats entre les méthodes peut provenir du fait que l'utilisation de plusieurs composantes patrimoniales, doit permettre une meilleure restitution de la distribution des patrimoines au delà du seuil de 450 000 €. Elle est peut-être également liée au fait que l'approche multivariée à mieux tenu compte de l'hétérogénéité que l'approche récapitulative ou que l'approche avec composantes indépendantes. L'approche récapitulative distinguait uniquement les ménages propriétaires de leur résidence principale des autres, tandis que dans l'approche multivariée, nous différencions 8 catégories de ménages, en fonction de la composition de leur portefeuille. Par ailleurs les modèles ne peuvent pas être vrais simultanément car dans un cas c'est le résidu du modèle pour le logarithme du patrimoine total qui suit une loi log-normale, alors que dans l'autre il s'agit du résidu du modèle pour le logarithme des composantes. Ce sont donc des modèles descriptifs même si nous nous sommes inspirés de modèles économiques lors de leur spécification.
La deuxième méthode détaillée dans l'Annexe 1 est basée sur des modèles univariés indépendants, néanmoins le conditionnement par les informations sur le patrimoine total, financier et l'ISF introduit de la dépendance ex-post. On peut justifier ce modèle par un mécanisme de choix de portefeuille. Les tests basés sur les résidus généralisés présentés en Annexe 2 ne permettent donc pas de tester l'indépendance du départ. Cependant, même en conditionnant par le niveau de patrimoine financier, le niveau de patrimoine total et le seuil d'imposition sur la fortune, certaines caractéristiques, comme les préférences pour le risque et pour le temps, qui déterminent les choix de portefeuille, ne sont pas observées, ou seulement pour une partie de l'échantillon. La présence de cette hétérogénéité inobservée introduit une corrélation potentielle entre les différentes composantes patrimoniales, même non conditionnellement au niveau de patrimoine total ce qui justifie en soit l'utilisation d'un modèle réellement multivarié.

Bibliographie

Arnold S.F. (1993), « Gibbs sampling », *Handbook of Statistics,* n°9, pp. 599-625.

Arrondel L., Masson A. et Verger D. (2004a), « Les comportements de l'épargnant à l'égard du risque et du temps », *Economie et Statistique*, n°374-375, pp. 9-19.

Arrondel L., Masson A. et Verger D. (2004b), « Préférences individuelles et disparités de patrimoine », *Economie et Statistique*, n°374-375, pp.129-157.

Cordier M., Houdré C. et Rougerie C. (2006), « Les inégalités de patrimoine des ménages entre 1992 et 2004 », *Insee Références - Les revenus et le patrimoine des ménages,* pp. 47-58.

Deville J.C., Särndal C.E. et Sautory O. (1993), « Generalized raking procedures in survey sampling», *Journal of the American Statistical Association,* n°88, pp. 1013-1020.

Dell F., d'Haultfoeuille X., Février P. et Massé E. (2002), « Mise en œuvre de calcul de variance par linéarisation » in *Actes des Journées de Méthodologie Statistique,* http://jms.insee.fr/site/index.php.

Gautier E. (2005), « Eléments sur les mécanismes de sélection dans les enquêtes et sur la non-réponse non-ignorable » in *Actes des Journées de Méthodologie Statistique,* http://jms.insee.fr/site/index.php.

Gourieroux C., Monfort A., Renault E. et Trognon A. (1987a), « Generalized residuals », *Journal of Econometrics,* n°34, pp. 5-32.

Gourieroux, C.; Monfort, A., Renault, E. et Trognon, A. (1987b), « Simulated residuals », *Journal of Econometrics,* n°34, pp. 201-252.

Heeringa S.G., Little R.J.A. et Raghunathan T.E. (2002), « Multivariate imputation of coarsened survey data on household wealth » in *Survey Nonresponse* edited by Robert M. Groves, et al., Wiley, pp. 357-372.

Juster T.F. et Smith J.P. (1997), « Improving the quality of economic data: lessons from the HRS and AHEAD », *Journal of the American Statistical Association*, n°92, pp. 1268-1278.

Little R.J.A. et Rubin D.B. (2002), « Statistical Analysis with Missing Data », 2nd edition. *Wiley.*

Lollivier S. et Verger D. (1989), « D'une variable discrète à une variable continue : la technique des résidus simulés » in *Mélanges économiques - Essais en l'honneur de Edmond Malinvaud,* Economica.

McCulloch R. et Rossi P.E. (1994), « An exact likelihood analysis of the multinomial probit model », *Journal of Econometrics*, n°64, pp. 207-240.

Meyn S.P. et Tweedie R.L. (1993), « Markov Chains and Stochastic Stability », Springer.

Robert C.P. (1995), « Simulation of truncated normal variables », *Statistics and Computing*, n°5, pp. 121-125.

Robert C.P. (2007), *The Bayesian Choice*, Springer.

Robert C.P. et Casella G. (2004), *Monte Carlo Statistical Methods*, 2nd edition, Springer.

Roberts G.O. et Polson N.G. (1994), « On the geometric convergence of the Gibbs sampler », *Journal of the Royal Statistical Society – B*, n°56.

Roberts G.O. et Smith A.F.M. (1994), « Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms », *Stochastic Processes and Their Applications*, n°49, pp. 207-216.

Schafer J.L. (2001), *Analysis of Incomplete Multivariate Data*, 2nd edition, Chapman & Hall.

Tierney L. (1994), « Markov chains for exploring posterior distributions », *The Annals of Statistics*, n°22, pp.1701-1728

Train K.E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press.

Annexes

Annexe 1
Compléments sur l'approche modèle à composantes indépendantes et prise en compte des contraintes mobilisant plusieurs composantes[25]

Pour chaque type de produit, le montant en intervalle est affiné grâce aux informations mobilisant plusieurs composantes. Des modèles linéaires Gaussiens pour le logarithme de la variable avec observations censurées sont estimés sur l'ensemble des ménages détenteurs.

Logements
La première composante des logements est la résidence principale. En plus des variables socio-démographiques, le modèle comprend en particulier le type d'habitation (individuel /collectif), le type d'agglomération et éventuellement la localisation géographique, ainsi que la surface habitable du logement et la surface au carré. Des modèles voisins sont estimés pour les autres logements (de jouissance ou de rapport).

Produits financiers
L'enquête recense 31 types de produits financiers dont une catégorie de « produits divers » qui regroupe les produits n'ayant pu être classés par le ménage ou l'enquêteur. Certains modèles sont estimés sur un échantillon de très petite taille et ne contiennent donc pour des problèmes d'identification et par souci de parcimonie que quelques caractéristiques socio-démographiques. Pour les produits plus courants, des caractéristiques liées au produit ont également pu être ajoutée comme l'année de souscription, le taux de rendement (pour les livrets d'épargne), le type de versement (pour les produits d'assurance vie), la part investie en actions (toujours pour les produits d'assurance-vie).

Patrimoine professionnel exploité
Pour les terres et terrains exploités par le ménage à titre professionnel, les modèles incluent la nature du terrain (vignes, terres labourables ou autre), la zone géographique et la surface agricole utile. Pour les autres biens professionnels, plusieurs catégories d'indépendants ont été distinguées (agriculteurs, professions libérales, commerçants et artisans).

Patrimoine professionnel non exploité
Très peu de variables produits sont disponibles dans l'enquête pour ces composantes.

Prise en compte du « reste » et de l'information sur l'ISF
Le "reste", rassemblant les composantes non décrites dans l'enquête, a été lui aussi modélisé. Nous avons pu reconstruire des intervalles censés contenir le montant de ce reste grâce aux informations mobilisant plusieurs composantes : la tranche du patrimoine total et l'imposition ou non à l'ISF. Pour ce qui est de la question récapitulative sur le patrimoine total incluant le reste, nous exploitons les intervalles pour chacun des produits décrits dans l'enquête et la tranche pour le patrimoine total $\left[PT_{bas}, PT_{haut}\right)$. Nous avons :

$$RESTE \in \left[PT_{bas} - \sum C^{i}_{haut}, PT_{haut} - \sum C^{i}_{bas}\right)$$

où C^{i}_{bas} et C^{i}_{haut} sont les bornes inférieures et supérieures des composantes. Ce calcul est relativement fruste. Les restes les plus importants se trouvent certainement dans le haut de la distribution (possession d'œuvres d'art, etc.). Cependant, ces ménages possèdent déjà beaucoup d'autres produits et le plancher de la dernière tranche de patrimoine globale étant relativement bas 450 000 €, la borne inférieure est donc très souvent 0 pour les patrimoines élevés et il n'y a pas de borne supérieure. Ces restes ne contribueraient pas alors la vraisemblance. Or, le "reste" est un mélange les biens durables et des biens de luxe (bijoux, œuvres d'art...) et on peut penser que les ménages ayant le moins de patrimoine aient essentiellement des biens durables. Les "restes" des ménages situés dans le haut de la distribution

[25] Approche utilisée en production afin d'effectuer l'imputation du fichier. La méthode multivariée a été développée en parallèle.

seraient alors simulés dans une loi estimée sur le bas de la distribution, c'est-à-dire sur la partie "biens durables". Voilà pourquoi nous avons également rétréci les fourchettes pour le "reste" à l'aide de l''appariement avec les données fournies par la Direction Générale des Impôts. Elles nous permettent de savoir si un ménage a été imposé sur la fortune en 2003 et en 2004[26]. L'enquête Patrimoine ne permet pas de calcul exact du patrimoine imposable, mais nous avons pu en calculer un majorant. Ce majorant est constitué de l'ensemble des biens, l'endettement étant déduit, et on déduit un abattement de 20% sur la valeur de la résidence principale. Le patrimoine exploité du ménage n'est pas comptabilisé, hormis pour des ménages possédant moins de 25% de leur entreprise. Une partie seulement du reste est imposable mais si nous souhaitons uniquement construire un majorant du reste nous pouvons l'inclure dans le calcul. Le seuil actuel de l'imposition sur la fortune est de 720 000 €. Utiliser l'information que le majorant du patrimoine imposable doit être supérieur à 720 000 € permet d'affiner les intervalles du reste pour les patrimoines les plus élevés. Ceci a permis de construire une borne inférieure pour la composante de reste pour à peu près un ménage imposé à l'ISF sur trois[27]. Afin de capter les spécificités de ce reste, l'estimation a été conduite séparément sur quatre groupes de ménages : les indépendants, les jeunes non-indépendants, les cadres-professions intermédiaires non indépendants, les ouvriers-employés non-indépendants.

La méthode mobilise, par rapport à la méthode basée exclusivement sur la variable synthétique, toute l'information disponible dans l'enquête. Cependant, les simulations du patrimoine global obtenue par sommation et incluant le reste ne sont pas entièrement satisfaisantes. On s'attend que le rajout de ce "reste" aux autres composantes de patrimoine fasse augmenter la part de patrimoine détenue par les ménages les plus aisés (prise en compte de variables de luxe, qui devraient se retrouver dans le haut de la distribution). Or nous observons le phénomène inverse. Remarquons que le phénomène inverse se passe avec l'approche multivariée où nous mesurons bien une augmentation conforme à l'intuition.

Tableau 4 : Estimateur de sondage d'indicateurs d'inégalités à partir de composantes simulées

	Simulation 1		**Simulation 2**		**Moyenne**	
	Ensemble des composantes et le « reste »	**Ensemble des composantes sans le « reste »**	**Ensemble des composantes et le « reste »**	**Ensemble des composantes sans le « reste »**	**Ensemble des composantes et le « reste »**	**Ensemble des composantes sans le « reste »**
Part du patrimoine détenu par...						
… les ménages les plus riches						
1%	*12,9*	*13,13*	*12,15*	*12,49*	*12,525*	*12,81*
10%	*45,03*	*45,82*	*44,51*	*45,45*	*44,77*	*45,63*
… les ménages les moins riches						
10%	*0,12*	*0,02*	*0,12*	*0,02*	*0,12*	*0,02*
50%	*8,71*	*7,31*	*8,84*	*7,37*	*8,775*	*7,34*
Gini	***0,6214***	***0,6411***	***0,6176***	***0,6385***	***0,6195***	***0,6398***
Moyenne (€)	*189 133*	*167 466*	*188 804*	*166 546*	*188 969*	*167 006*
Médiane (€)	*113 154*	*99 347*	*113 390*	*99 640*	*113 272*	*99 494*
d9/d5	*3,72*	*3,89*	*3,73*	*3,83*	*3,725*	*3,86*
moyenne/d5	*1,67*	*1,69*	*1,67*	*1,68*	*1,67*	*1,685*
Theil	***0,77***	***0,82***	***0,74***	***0,8***	***0,755***	***0,81***

Source: Enquête Patrimoine 2004, calculs des auteurs

[26] Nous avons utilisé les déclarations pour 2004, car il s'agit du patrimoine possédé par le ménage en janvier 2004, c'est à dire la date la plus proche de la collecte effectué (celle-ci s'est faite entre octobre 2003 et janvier 2004).

[27] On peut regretter de ne pas avoir utilisé dans cette méthode l'information qui nous est fournie lorsqu'un message n'est pas imposé à l'ISF. Afin d'avoir un traitement symétrique et pour éviter l'introduction d'un biais par asymétrie, nous avons tenu compte de cette seconde information dans la méthode multivariée.

Compte tenu du temps de calcul lié à l'inefficacité[28] de l'acceptation/rejet pour simuler des lois tronquées en dimension élevée lorsqu'on prend comme loi instrumentale la loi non tronquée, nous avons effectué peu de simulations. Cependant, sur les quelques simulations faites, les résultats semblent moins volatiles d'une simulation à l'autre, que pour la méthode de simulation du patrimoine global.

[28] On aurait pu utilement utiliser un échantillonnage de Gibbs, cf. Robert (1995), ou procéder par échantillonnage préférentiel et utiliser le simulateur GHK.

Annexe 2
Tester la dépendance avec les résidus généralisés

Le cadre dans lequel Lollivier et Verger (1988) effectuent leurs imputations est univarié. Chaque composante patrimoniale est imputée indépendamment des autres. Compte-tenu que le but, dans cet article, est d'estimer un indicateur d'inégalité calculé sur le patrimoine total (la somme des composantes), cette hypothèse est sûrement très forte. Il est d'ailleurs possible de tester l'indépendance des composantes deux à deux en s'appuyant sur le calcul des « résidus généralisés » présenté dans Gouriéroux et al. (1987a). Il est remarquable que cette approche ne repose pas sur la spécification d'un modèle joint. On teste la corrélation des composantes au niveau de la détention, en calculant les résidus généralisés dans le cadre de modèle *Probit*, ainsi qu'au niveau des montants de chaque composante, à détention donnée, en calculant les résidus généralisés dans le cadre de modèle de régression censurée. Les tests porteront sur la corrélation de composantes agrégées, telles que celles utilisées dans notre approche Bayésienne, ainsi qu'un à niveau plus fin, notamment pour le patrimoine financier.

Encadré 5

Résidus généralisés

Le principe du test de corrélation de deux variables latentes à partir des résidus généralisés est présenté d'après Gouriéroux et al. (1987a). Le modèle bivarié s'écrit de manière assez générale comme suit:

$$\begin{array}{rcl} y_{1i}^* & = & m_1(x_i,\alpha_1)+w_{1i} \\ y_{2i}^* & = & m_2(x_i,\alpha_2)+w_{2i} \end{array} \quad \text{avec} \quad \begin{pmatrix} w_{1i} \\ w_{2i} \end{pmatrix} \rightsquigarrow \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

On note θ le vecteur de l'ensemble des paramètres. Dans ce modèle latent (les variables sont observées de manière imparfaite, ex. censurées), la composante du score associé au coefficient de corrélation σ_{12} est proportionnelle à $\sum_{i=1}^n w_{1i}w_{2i}$ et donc à $\sum_{i=1}^n E(w_{1i}w_{2i}|y_i)$ dans le modèle observé, où y_i correspond aux observables. En notant $\hat{\theta}_n^0$ l'estimateur du maximum de vraisemblance de θ sous l'hypothèse nulle d'indépendance des variables on a :

$$\begin{array}{rcl} \sum_{i=1}^n \hat{\mathbb{E}}^0(w_{1i}w_{2i}|y_i) & = & \sum_{i=1}^n \hat{\mathbb{E}}^0(w_{1i}|y_i)\hat{\mathbb{E}}^0(w_{2i}|y_i) \\ & = & \sum_{i=1}^n \hat{\tilde{w}}_{1i}^0\hat{\tilde{w}}_{2i}^0 \end{array}$$

où $\hat{\tilde{w}}_{1i}^0$ et $\hat{\tilde{w}}_{2i}^0$ sont les résidus généralisés associés à chaque équation. Gouriéroux et al montrent que le test du score de l'hypothèse nulle s'appuie sur la fonction des résidus généralisés suivantes :

$$\xi_n = \frac{\left(\sum_{i=1}^n \hat{\tilde{w}}_{1i}^0\hat{\tilde{w}}_{2i}^0\right)^2}{\sum_{i=1}^n (\hat{\tilde{w}}_{1i}^0\hat{\tilde{w}}_{2i}^0)^2}$$

Dans le cas d'un modèle *Probit* bivarié, par exemple pour la détention jointe de deux composantes patrimoniales, les résidus généralisés s'écrivent :

$$\hat{\tilde{w}}_{ki}^0 = \frac{\varphi(x_i'\hat{\alpha}_k^0)}{\Phi(x_i'\hat{\alpha}_k^0)\cdot\left(1-\Phi(x_i'\hat{\alpha}_k^0)\right)}\cdot\left(y_i-\Phi(x_i'\hat{\alpha}_k^0)\right)$$

Pour aller plus loin, on modélise conjointement détention et encours, en supposant, par souci de simplification, que les détenteurs d'un produit y ont investi une somme supérieure à 1 euro. On observe alors la détention du produit k, et, pour les détenteurs, une borne inférieure $\underline{y}_{ki}$ et une borne supérieure $\overline{y}_{ki}$. En passant au log, le modèle latent s'écrit sous la forme suivante :

$$y_{ki}^* = x_i'\cdot\theta+\sigma\cdot u_i$$

avec

$$
\begin{array}{rcccl}
 & & y_{ki}^{*} & < & 0 \quad \text{non détention} \\
\underline{y}_{ki} & < & y_{ki}^{*} & < & \overline{y}_{ki} \quad \text{détention, aucune borne manquante} \\
0 & < & y_{ki}^{*} & < & \overline{y}_{ki} \quad \text{détention, borne inférieure manquante} \\
\underline{y}_{ki} & < & y_{ki}^{*} & < & +\infty \quad \text{détention, borne supérieure manquante} \\
0 & < & y_{ki}^{*} & < & +\infty \quad \text{détention, bornes inérieure et supérieure manquantes}
\end{array}
$$

Dans ce cas, les résidus généralisés s'écrivent :

$$
-\frac{\varphi\left(\frac{\overline{y}_i - x_i'\theta}{\sigma}\right) - \varphi\left(\frac{\underline{y}_i - x_i'\theta}{\sigma}\right)}{\Phi\left(\frac{\overline{y}_i - x_i'\theta}{\sigma}\right) - \Phi\left(\frac{\underline{y}_i - x_i'\theta}{\sigma}\right)}
$$

Dépendance dans la détention

On teste l'indépendance deux à deux des composantes, conditionnellement aux variables explicatives x. A un niveau agrégé, c'est-à-dire pour les 5 composantes retenues dans le modèle multivarié, l'indépendance est dans la plupart des cas rejetée (tableau ?). Seul le « reste », composé des biens durables, bijoux et objets d'art semble être indépendants des autres composantes de patrimoine, en dehors de la résidence principale.

Tableau 5 : P-value pour le test d'indépendance des variables latentes de détention deux à deux fondé sur le score

	Résidence principale	Autres logements	Patrimoine professionnel	"Reste"
Patrimoine financier	1,16E-08	1,22E-04	6,72E-04	0,19
résidence principale		2,17E-04	5,38E-05	6,51E-07
autres logements			9,38E-04	0,14
Patrimoine professionnel				0,94

Source : Enquête Patrimoine 2004, calculs des auteurs

A un niveau plus fin, la détention de tel ou tel actif est également bien souvent corrélée à la détention de tel autre (Tableau 6).

Tableau 6 : P-value pour le test d'indépendance des variables latentes de détention deux à deux fondé sur le score

		P-value
Sicav ou FCP	Obligations	1,28E-12
	Pea	0
	Compte-titres	0
Livret Jeune	Codevi	0,44
	Livret A ou Bleu	1,00E-03
Biens professionnels	Actions cotées	1,95E-05
	Actions non cotées	0

Source : Enquête Patrimoine 2004, calculs des auteurs

Pour le patrimoine financier en particulier, la présence de Sicav ou de FCP dans le patrimoine n'est pas indépendante de celle d'obligations ou de la détention d'un PEA par exemple. En revanche, on peut considérer que la détention d'un Codevi n'est pas corrélée à celle d'un livret Jeune.

Dépendance des encours

La modélisation de la détention et des encours sous la forme d'un modèle Tobit censuré et le test de corrélation entre composantes deux à deux conduit la plupart du temps à rejeter l'indépendance des composantes financières (tableau 3). Le montant détenu en Sicav ou en FCP n'est pas indépendant de l'épargne placée sous forme d'obligations ou dans un PEA. En revanche, l'épargne sur les livrets jeunes n'est pas corrélée à celle placée sur un Codevi.

Tableau 7 : P-value associée au test d'indépendance des encours deux à deux fondé sur le score

		P-value
Sicav ou FCP	Obligations	4,06E-11
	PEA	0
	Compte-titres	0
Livret Jeune	Codevi	0,22
PEA	Actions non cotées	0

Source : Enquête Patrimoine 2004, calculs des auteurs

Au vu de l'ensemble de ces résultats, il semble important de chercher à spécifier un modèle multivarié pour les différentes composantes patrimoniales.