Université de Montréal

Gérer le risque d'échantillonnage en économétrie financière : modélisation et contrôle

par

Bertille Antoine

Département de sciences économiques

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Philosophiae Doctor (Ph.D.)

en sciences économiques, option économétrie et économie financière

Août 2007

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Gérer le risque d'échantillonnage en économétrie financière : modélisation et contrôle

présentée par

Bertille Antoine

a été évaluée par un jury composé des personnes suivantes :

Président-rapporteur	Marine Carrasco
Directeur de recherche	Eric Renault
Examinateur externe	Alastair Hall-Univ. of Manchester Oxford Road
Membre du jury	Marc Henry
Représentant du doyen de la FES	Christian Léger

Sommaire

L'objet de cette thèse est le traitement de contextes d'application, en particulier dans le domaine de l'économie financière, où le point de vue asymptotique traditionnel peut être trompeur. Chaque essai propose alors une méthode pour affiner les approximations asymptotiques en présence d'échantillons d'observations qui, en pratique, sont toujours finis.

Le premier essai se place dans la lignée de la littérature récente sur les instruments faibles. Nous adaptons le contexte général de la méthode des moments généralisée (GMM) afin de lier plus spécifiquement la faible identification aux instruments, c'est-à-dire aux conditions de moment. Ainsi, contrastant avec la plupart des méthodes existantes, la partition d'intérêt entre les paramètres structurels fortement et faiblement identifiés n'est pas spécifiée a priori : elle s'obtient plutôt après une rotation dans l'espace des paramètres. Par ailleurs, nous nous concentrons ici sur le cas d'identification presque-faible pour lequel la déficience de rang est atteinte à la limite à un taux de convergence plus lent que l'usuel racine-T. Dans ce contexte, les estimateurs GMM de tous les paramètres convergent, à des taux possiblement plus lents que d'habitude. Cela nous permet de valider les approches de test standard comme Wald ou GMM-LM. De plus, nous identifions et estimons des directions dans l'espace des paramètres pour lesquelles la convergence au taux racine-T est maintenue. Ces résultats sont d'un intérêt direct pour les applications pratiques, et ce, sans que la connaissance ou l'estimation du taux de convergence plus lent ne soit requise. Nous proposons des illustrations Monte-Carlo pour deux modèles économétriques : le modèle de régression linéaire avec variables instrumentales à une équation et le modèle d'évaluation d'actifs CAPM avec consommation.

Le deuxième essai complète le premier en réalisant une étude comparative de puissance pour deux tests de la littérature GMM avec instruments (presque)-faibles : le test de score classique, valide dans le cadre du premier essai, et le test de Kleibergen ou score modifié. Plus

généralement, nous comparons deux approches : d'une part, à l'image du premier essai, spécifier les problèmes d'identification, via le comportement des conditions de moment, permet d'appliquer les procédures de test standard ; d'autre part, comme dans Kleibergen (2005), ne pas préciser le cadre d'identification requiert une modification de la statistique du score.

Dans le troisième essai, nous proposons une nouvelle méthode d'inférence, la procédure Modified-Wald, afin de pallier au mauvais comportement (connu) des tests de Wald lorsque l'identification n'est plus assurée à la frontière de l'espace des paramètres. Nous nous concentrons ici sur le ratio de paramètres multidimensionnel lorsque le dénominateur est proche de la singularité. Notre méthode est basée sur la statistique de Wald : le contenu informationnel de l'hypothèse nulle d'intérêt est intégré dans le calcul de sa métrique. Cette correction préserve la tractabilité de la méthode et permet d'obtenir une région de confiance non bornée lorsque nécessaire. La procédure de Wald standard produit habituellement une région de confiance bornée : celle-ci est invalide pour toute taille d'échantillon donnée dans la mesure où sa probabilité de couverture est nulle. La seule manière de remédier à ce problème est d'obtenir des régions de confiance non bornées avec une probabilité non nulle. Une simulation compare les propriétés d'inférence des procédures Wald et Modified-Wald avec un ratio bidimensionnel. Nous considérons aussi le modèle de régression linéaire avec variables instrumentales à une équation lorsque les propriétés identifiantes des instruments varient.

Pour finir, contrastant avec les trois premiers essais qui restent dans le domaine de la théorie statistique asymptotique, le quatrième essai adopte un point de vue décisionnel dans le domaine du choix de portefeuille. Un défi important associé à l'allocation de portefeuille se produit lorsque les caractéristiques (inconnues) de la distribution des rendements financiers sont remplacés par des estimés. Cela introduit du risque dit d'estimation, crucial pour la gestion de portefeuille, au même titre que le risque financier traditionnel. Cet essai se concentre sur une nouvelle mesure de performance par rapport à la littérature existante. Nous empruntons aux praticiens et évaluons les différentes allocations de fonds à travers leur vraisemblance à battre un niveau de référence donné. Ensuite, le portefeuille optimal, qui incorpore alors le risque d'estimation, est connu explicitement et ne dépend d'aucun paramètre de nuisance. Une étude de Monte-Carlo simple compare plusieurs stratégies d'investissement de la littérature.

Mots clés : GMM ; variables instrumentales ; identification (presque)-faible ; test K ; test du score ; ratio de paramètres ; Wald ; région de confiance non bornée ; théorie du portefeuille ; risque d'estimation ; performance de référence ; efficacité moyenne-variance.

Summary

The objective of this thesis is to study designs, particularly in the field of financial economics, where the asymptotic point of view may be misleading. Each essay proposes a method to refine the asymptotic approximations in the presence of samples which are, in practice, always finite.

The first essay is in line with the recent literature on weak instruments. We propose to adapt the general framework of the Generalized Method of Moments (GMM) in order to specifically relate weakness to the instruments, that is the moment conditions. As a consequence, in contrast with most of the existing literature, the relevant partition between strongly and weakly identified structural parameters is not specified a priori but rather achieved after a well-suited rotation in the parameter space. In addition, we focus here on the case dubbed nearly-weak identification where the drifting DGP introduces a limit rank deficiency reached at a rate slower than root-T. This framework ensures that the GMM estimators of all parameters are consistent but at rates which may be slower than usual. This allows us to verify the validity of the standard testing approaches like Wald and GMM-LM tests. Moreover, we identify and estimate directions in the parameter space where root-T convergence is maintained. These results are all directly relevant for practical applications without requiring knowledge or estimation of the slower rate of convergence. We provide Monte-Carlo illustrations for two econometric models: the single-equation linear IV model and the consumption based CAPM.

The second essay completes the first one with a comparative study of the power of two tests proposed within the GMM literature when the identification is (nearly)-weak: the standard score test, valid in the framework of chapter 1, and the K-test or modified score test. In a more general sense, we are comparing two approaches with respect to identification issues: on one hand, as shown in the first essay, specifying identification issues through moment

conditions allows the application of standard test procedures; on the other hand, as shown by Kleibergen (2005), in the absence of identification issue specification a modification of the score test statistic is required.

In the third essay, we propose a new inference method, the Modified-Wald procedure, to overcome some issues of the well-documented poor behavior of Wald tests when identification is lost at the frontier of the parameter space. We focus here on the multidimensional ratio of parameters when the denominator is close to singularity. This method is based on the Wald statistic. The key idea consists of integrating the informational content of the null hypothesis of interest in the computation of its metric. This correction, while preserving the computational tractability of the method, allows for unbounded confidence regions when needed. A standard Wald test usually provides a bounded confidence region: this region is invalid for any given sample size in the sense that its coverage probability is zero. The only way to surmount this issue is to write confidence regions with a nonzero probability of being unbounded. A simulation exercise compares the inference properties of the Wald and Modified-Wald procedures with a bidimensional ratio. We also consider the single-equation linear IV model in cases where the identifying properties of the instruments may vary.

Finally, in contrast to the first three essays which remain in the framework of statistical asymptotic theory, the fourth essay adopts a decisional point of view in portfolio choice. An important challenge in portfolio allocation arises when the (true) characteristics of returns distribution are replaced by estimates. This introduces estimation risk, which is a crucial component of portfolio management, just like the traditional financial risk. This essay differs from existing literature by virtue of its focus on a different measure of performance. We borrow from practitioners and evaluate the funds allocations based on their likelihood of beating a benchmark. Then, the optimal portfolio which accounts for estimation risk is known in closed-form and does not depend on any nuisance parameter. This investment rule corresponds to a mean-variance investor with a corrected, sample-dependent risk aversion. A simple Monte-Carlo study involving five risky assets is used to compare several investment strategies.

Key Words: GMM; Instrumental variables; (Nearly)-weak identification; K-test; Score test; Parameter ratio; Wald; Unbounded confidence region; Weak instruments; Portfolio theory; Estimation risk; Benchmark performance; Mean-variance efficiency.

Table des matières

Liste des tableaux

Table des figures

Remerciements

Je tiens tout d'abord à remercier mon directeur de recherche, Éric Renault. Il a été présent à chaque étape du processus : depuis mes premières hésitations à m'engager (à long terme) en recherche lors de mon arrivée à Montréal, jusqu'au dernier rebondissement de cette histoire qui me conduit à présent à Vancouver... Sans lui, rien de tout ceci ne se serait passé ainsi. Un énorme merci pour les multiples facettes de son soutien : méthodologique, théorique, moral et financier. Je suis honorée d'avoir pu évoluer à ses côtés.

Je remercie les laboratoires de recherche et les institutions qui m'ont accueillie et financée : le Centre Interuniversitaire de Recherche en Économie Quantitative (CIREQ) et le département d'économie de l'Université de Montréal, le Centre Interuniversitaire de Recherche en Analyse des Organisations (CIRANO), les départements d'économie et de statistique de UNC-Chapel Hill. Cette thèse a également bénéficiée du financement de bourses de recherche de l'Institut de Finance Mathématique (IFM2) de Montréal et du Conseil de Recherche en Sciences Humaines du Canada (CRSH).

Je remercie les participants aux conférences, séminaires, groupes de travail... pour leurs discussions et commentaires.

Enfin, je remercie toutes les personnes qui m'ont épaulée quand j'en avais besoin : parents, amis, collègues... la liste est longue. Merci pour votre présence et vos encouragements : c'est aussi grâce à vous si j'y suis arrivé.

Introduction générale

Fournir de l'inférence de qualité sur les paramètres d'intérêt a toujours été une question centrale en économétrie. Pour ce faire, l'approche fréquentiste se base sur deux résultats essentiels : la loi des grands nombres et le théorème de la limite centrale (TLC). Ils assurent respectivement que les vraies valeurs (inconnues) des paramètres sont connues asymptotiquement, c'est-à-dire quand la taille de l'échantillon observé tend vers l'infini, et approchées par des estimateurs asymptotiquement gaussiens. Sous des hypothèses de régularité standard, il est communément admis que les résultats précédents sont vérifiés. Dans ces conditions, l'inférence à partir d'une statistique de Wald est très prisée par les praticiens : on calcule un estimateur de la quantité d'intérêt et son comportement asymptotique est fournit par le TLC ; s'en suivent alors les tests et régions de confiance associés. Ces dernières sont construites, par exemple, en inversant cette statistique de Wald : cela signifie simplement que les valeurs des paramètres pour lesquelles le test n'est pas significatif sont collectées. De telles régions sont généralement bornées.

Plus récemment, un intérêt particulier s'est fait ressentir pour fournir de l'inférence valide lorsque l'identification des paramètres n'est plus complètement assurée. Deux situations peuvent entraîner une perte partielle ou totale de l'identification : soit, l'identification est tout simplement perdue à la frontière de l'espace des paramètres ; soit, les conditions qui assurent l'identification du modèle font défaut. Dans le premier cas, il est facile d'imaginer une transformation des paramètres qui ne serait valide que dans un sous-ensemble de l'espace des paramètres d'origine : par exemple, un ratio n'est défini que lorsque le dénominateur est non nul. Dans le second cas, on peut penser à l'un des cheval de bataille de la recherche empirique en économie, à savoir l'instrumentation des variables exogènes. Plus précisément, un modèle structurel fait généralement intervenir des variables explicatives endogènes, c'est-à-dire liées

au terme d'erreur. Ceci invalide l'utilisation de la méthode des moindres carrés ordinaires et l'on a recours à des variables instrumentales (IV) ou instruments pour assurer l'identification des paramètres du modèle et mener à bien l'inférence statistique. Les instruments sont des variables auxiliaires exogènes, ou encore non corrélées avec le terme d'erreur, qui doivent être suffisamment pertinentes : c'est-à-dire suffisamment bien corrélées avec les variables explicatives endogènes. Lorsque cette corrélation est faible, l'identification des paramètres n'est plus complètement assurée.

La perte partielle ou totale d'identification peut entraîner des comportements asymptotiques inhabituels chez certaines statistiques de test. Plus généralement, les méthodes d'inférence standard peuvent être invalidées. Plusieurs articles ont documenté la faible performance des méthodes et approximations asymptotiques usuelles : entre autres, Nelson et Startz (1990), Bound, Jaeger et Baker (1995) et Staiger et Stock (1997). Plusieurs pistes de recherche ont alors été envisagées dans la littérature pour fournir des méthodes d'inférence fiables.

L'économètre peut d'abord envisager une modification du cadre de travail en changeant le scénario asymptotique, afin de pouvoir dériver le comportement asymptotique des statistiques de test considérées. En d'autres termes, les propriétés d'identification du modèle sont maintenant liées artificiellement à la taille de l'échantillon. Par exemple, dans le cadre d'un modèle structurel linéaire à équations simultanées, Staiger et Stock (1997) modélisent la corrélation entre les instruments et les variables endogènes comme inversement proportionnelle à la taille de l'échantillon à la puissance 1/2 : cette situation est connue sous le nom d'identification faible. Plus récemment, Hahn et Kuersteiner (2002) considèrent différentes puissances de la taille de l'échantillon qui caractérisent le degré d'identification : par exemple, l'identification est presque-faible lorsque la puissance est strictement comprise entre 0 et 1/2.

Une autre approche consiste à modifier directement les statistiques de test existantes afin de les rendre robustes aux différents cas d'identification. Par exemple, dans le cadre de la méthode des moments généralisée (GMM), Kleibergen (2005) propose le test K ou test du score modifié : l'estimateur usuel du jacobien espéré est remplacé par un estimateur qui est asymptotiquement non corrélé avec la moyenne empirique des conditions de moment. Cette modification rend le test robuste aux instruments faibles.

Enfin, une dernière voie majeure de recherche se concentre sur des méthodes d'inférence

dites exactes. Elles ne s'appuient ni sur une hypothèse d'identification, ni sur la normalité asymptotique des estimateurs mais plutôt sur des statistiques pivotales robustes aux problèmes d'identification. Citons la première d'entre elles, la statistique de Anderson et Rubin ou statistique AR (Anderson et Rubin (1949)). Une démarche statistique classique consiste alors à dériver un système d'inférence à partir d'une statistique pivotale. Toute la difficulté réside dans l'obtention de tels pivots.

Les quatre essais de cette thèse traitent de contextes d'application, en particulier dans le domaine de l'économie financière, où le point de vue asymptotique traditionnel peut être trompeur. Chaque essai propose alors une méthode pour affiner les approximations asymptotiques en présence d'échantillons d'observations qui, en pratique, sont toujours finis.

Le premier essai se concentre sur les problèmes d'identification liés à des instruments presque-faibles. Notre approche consiste à adapter le contexte général de la méthode des moments généralisée (GMM) afin que la faiblesse des instruments soit en lien direct avec les conditions de moment. Plus précisément, ces dernières sont partitionnées suivant l'information statistique qu'elles véhiculent : un groupe de conditions de moment standard associé au taux de convergence habituel et un groupe faible associé à un taux plus lent. Les paramètres structurels sont alors estimés de manière usuelle, mais à des taux de convergence possiblement plus lents. C'est le cas, en particulier, lorsque le paramètre d'intérêt représente une caractéristique fine de la population qui n'est que faiblement identifiée par les observations à notre disposition : par exemple, la caractérisation des événements rares, le prix des actifs contingents à de tels événements ou encore le niveau des primes associées à des risques à peine prévisibles.

Le deuxième essai complète le premier en réalisant une étude comparative de puissance pour deux tests proposés dans la littérature GMM avec instruments (presque)-faibles : le test de score classique, valide dans le cadre du premier essai, et le test de Kleibergen (2005) ou score modifié.

L'approche développée dans le troisième essai est plus spécifiquement adaptée au cas où le défaut d'identification du paramètre d'intérêt n'apparaît qu'à la frontière du domaine autorisé des paramètres. Elle considère des régions de confiance potentiellement non bornées dans certaines configurations des données d'observation. On ne devrait pas être surpris d'obtenir des régions non bornées lorsque qu'un paramètre n'est pas ou peu identifié : en effet, celles-

ci doivent simplement être interprétées comme un manque d'information disponible dans l'échantillon pour fournir de l'inférence précise sur ce paramètre.

Enfin, contrastant avec les trois premiers essais qui restent dans le cadre de la théorie statistique asymptotique, le quatrième essai adopte plus explicitement un point de vue décisionnel dans le contexte du choix de portefeuille. Le risque d'erreur statistique présent dans les moments estimés est ici considéré simultanément avec le risque financier, provenant de l'aléa des rendements : ceci, dans le but de proposer une gestion intégrée de ces deux risques. Toutefois, notre solution passe encore par une approche en termes de test statistique et peut donc être reliée, en ce sens, à la problématique générale de la thèse.

La contribution détaillée de ces quatre essais est à présent développée.

Le premier essai est basé sur un article rédigé conjointement avec Éric Renault. Dans cet essai, nous revisitons l'approche d'identification partielle développée par Phillips (1989), tout en maintenant l'identification complète de tous les paramètres, mais à des taux potentiellement plus lents. Nous conservons la normalité asymptotique des estimateurs GMM, déduite de l'identification de premier ordre ; cependant, le jacobien espéré peut disparaître lorsque la taille de l'échantillon augmente. À cet égard, nous sommes dans la lignée de la littérature récente sur les instruments faibles, qui, suivant l'approche pionnière de Staiger et Stock (1997) et de Stock et Wright (2000), capture l'identification faible à partir de conditions de moment empiriques. Toutefois, nous ne spécifions pas a priori le degré d'identification (fort ou faible) des paramètres. Nous considérons que la faiblesse doit être liée plus spécifiquement aux instruments, c'est-à-dire aux conditions de moment qui leur sont associées. Ainsi, la partition fort/faible des paramètres structurels ne peut être atteinte qu'après une rotation adéquate dans l'espace des paramètres.

Par ailleurs, tout comme Caner (2005), nous nous concentrons sur l'identification presque-faible dans laquelle la déficience de rang apparaît à la limite, à un taux plus lent que racine-T. De cette façon, tous les paramètres sont estimés de manière convergente, mais à des taux possiblement plus lents que d'habitude. Il est à noter que la déficience de rang asymptotique considérée garantit toujours des taux de convergence au moins égaux à $T^{1/4}$ pour tous les estimateurs GMM. C'est un contraste important avec l'approche de Stock et Wright (2000) :

en considérant une déficience asymptotique de rang atteinte au taux racine-T, les estimateurs GMM ne sont même pas convergents. Obtenir des estimateurs GMM convergents avec des taux bien définis (même s'ils sont potentiellement plus lents que la normale) nous permet de valider les approches de test standard comme Wald ou GMM-LM de Newey et West (1987). Par rapport à Kleibergen (2005), nous n'avons pas besoin de modifier les formules standard pour le test LM.

Il est évident que notre approche ne vise pas à capturer des cas sévères d'identification faible qui se produisent même lorsque la taille de l'échantillon est très grande (voir Angrist et Krueger (1991)). Toutefois, elle fournit au praticien des procédures d'estimation et d'inférence qui sont valides avec les formules standard, tout en l'avertissant que, dans certaines directions, les taux de convergence peuvent être plus lents que l'usuel racine-T. Ces résultats sont appliqués à un modèle d'équilibre général basé sur le modèle d'évaluation d'actifs CAPM avec consommation.

Le deuxième essai est basé sur un article rédigé conjointement avec Éric Renault. Il complète le premier essai en réalisant une étude comparative de puissance pour deux tests proposés dans la littérature GMM avec instruments (presque)-faibles : le test de score classique, valide dans le cadre du premier essai, et le test de Kleibergen ou score modifié. Plus généralement, il s'agit aussi de comparer deux approches : d'une part, à l'image du premier essai, la spécification des problèmes d'identification, via le comportement des conditions de moment, offre accès aux procédures de test standard ; d'autre part, comme dans Kleibergen (2005), ne faire aucune précision du cadre d'identification requiert une modification de la statistique du score.

Dans le troisième essai, nous considérons le ratio de paramètres multidimensionnel lorsque le dénominateur est proche de la singularité. Nous proposons une nouvelle méthode d'inférence, la procédure Modified-Wald. Cette méthode est basée sur la statistique de Wald : il s'agit d'intégrer le contenu informationnel de l'hypothèse nulle d'intérêt dans le calcul de sa métrique. Cette correction, tout en préservant la commodité des calculs, permet l'obtention de régions de confiance non bornées lorsque l'identification n'est plus complètement assurée. Le caractère borné des régions de confiance s'est révélé problématique depuis Dufour (1997). Dans le contexte de la quasi-identification locale (local almost identification), Dufour (1997) fournit des résultats sur la caractérisation des régions de confiance : sous certaines conditions

de régularité, ces régions doivent être non bornées avec une probabilité non nulle. En particulier, lorsque l'identification fait défaut, la plupart des ensembles de confiance de type Wald ont un niveau de confiance nul car ils sont presque sûrement bornés. En comparaison, notre procédure Modified-Wald, aussi attractive du point de vue computationnel, offre la possibilité d'obtenir des régions de confiance non bornées si nécessaire.

Par ailleurs, lorsque l'identification fait défaut à la frontière de l'espace des paramètres (dans l'esprit de Dufour (1997)), nous montrons que la probabilité d'obtenir une région de confiance non bornée atteint la borne supérieure de Dufour (1997). Lorsque les problèmes d'identification sont (artificiellement) reliés à la taille de l'échantillon (dans l'esprit du Pitman drift), cette probabilité dépend du taux de convergence vers la non-identification. Par exemple, avec une identification faible (taux égal à racine-T), cette probabilité est non-nulle mais plus petite que la borne supérieure précédente. Un exercice de simulation confirme les bonnes propriétés d'inférence de la procédure Modified-Wald par rapport à Wald avec un ratio bidimensionnel.

Dans le contexte du choix de portefeuille, un défi important intervient lorsque les caractéristiques (inconnues) de la distribution des rendements financiers sont remplacées par des estimés. Ce problème combine donc des difficultés d'ordre statistique à un problème d'économie financière classique consistant à choisir l'allocation de fonds optimale. Dans le quatrième essai, nous adoptons un point de vue décisionnel afin de développer une règle d'investissement qui incorpore à la fois le risque financier traditionnel et le risque d'estimation. Ce dernier provient directement du fait que, en pratique, les échantillons sont toujours de taille finie : ainsi, les estimés sont-ils toujours différents de leurs vraies valeurs respectives.

Pour répondre à cette question, nous nous concentrons sur une mesure de performance différente de la littérature. Nous empruntons aux praticiens et évaluons les différentes allocations de fonds à travers leur vraisemblance à battre un niveau de référence. Notre objectif est donc plus conservateur qu'une maximisation directe de la performance espérée du portefeuille (voir entre autres Markowitz (1959), Kan et Zhou (2006)). Toutefois, il correspond à l'intérêt direct de plusieurs industries : par exemple, les fonds de pension se doivent de garantir un niveau minimal de performance à leurs investisseurs. Pour un niveau de référence donné, nous déduisons une règle d'investissement explicite qui incorpore naturellement le risque d'estimation de la moyenne et ne dépend d'aucun paramètre de nuisance. Ainsi, elle

est directement applicable, sans recourir à aucune étape préalable sous-optimale.

Plus précisément, notre méthode de sélection de portefeuille se base sur un test unilatéral qui assure que la performance du portefeuille est au-dessus d'un niveau de référence donné ; ensuite, l'allocation optimale s'obtient en maximisant la p-valeur associée à ce test. C'est donc en combinant un outil statistique naturel et valide pour comparer des quantités aléatoires (ici les performances estimées des portefeuilles) à une mesure de performance directement construite à partir des intérêts des praticiens que nous proposons une règle d'investissement explicite qui intègre directement l'incertitude du problème.

Une étude Monte-Carlo simple, calibrée à partir de rendements mensuels des indices de stock pour les pays du G5, révèle le bon comportement de notre règle d'investissement en termes de performance espérée hors-échantillon et de stabilité dans le temps par rapport à d'autres règles de la littérature.

Chapitre I

Efficient GMM with Nearly-Weak Identification[†]

[†] This chapter is based on a paper co-authored with Eric Renault

1 Introduction

The cornerstone of GMM estimation is a set of population moment conditions, often deduced from a structural econometric model. The limit distributions of GMM estimators are derived under a central limit theorem for the moment conditions and a full rank assumption of the expected Jacobian. The latter assumption is not implied by economic theory and many circumstances where it is rather unjustified have been documented in the literature (see Andrews and Stock (2005) for a recent survey).

Earlier work on the properties of GMM-based estimation and inference in the context of rank condition failures includes Phillips (1989) and Sargan (1983). In the context of a classical linear simultaneous equations model, Phillips (1989) considers the case of a *partially identified* structural equation. He notes that, in case of rank condition failure, it is always possible to rotate coordinates in order to isolate estimable linear combinations of the structural parameters while the remaining directions are completely unidentified. Asymptotic theory of standard IV estimators in this context is then developed through the general framework of limited mixed Gaussian family. This approach of *partially identified* models differs from Sargan (1983) *first order under-identification*. While for the former there is nothing between estimable parameters with standard root-T consistent estimators and completely unidentified parameters, the latter considers that asymptotic identification is still guaranteed but it only comes from higher order terms in the Taylor expansion of first order optimality conditions of GMM: higher order terms become crucial when first order terms vanish. They are responsible for slower rates of convergence of GMM estimators like $T^{1/4}$ and may lead to non-normal asymptotic distributions like a Cauchy distribution or a mixture of normal distributions.

Our contribution in this essay is to revisit an approach of partial identification *à la* Phillips (1989), while maintaining, like Sargan (1983), the complete identification of all parameters, but at possibly slower rates. Moreover, we remain true to asymptotic normality of GMM estimators deduced from first order identification but with an expected Jacobian that may vanish when the sample size increases. In this respect, we are in the line of the recent literature on weak instruments, which, following the seminal approach of Staiger and Stock (1997) and Stock and Wright (2000), captures weak identification by drifting population moment conditions. With respect to the existing literature, the contribution of this essay is as follows.
First, in sharp contrast with most of the recent literature on weak instruments, we do not spec-

ify a priori which parameters are strongly or weakly identified. Conforming to the common wisdom that weakness should rather be assigned to specific instruments or more generally to some specific moment conditions, we follow Phillips (1989) to consider that the relevant partition of the set of structural parameters between strongly and weakly identified ones can only be achieved after a well-suited rotation in the parameter space. In nonlinear settings, this change of basis depends on unknown structural parameters and must itself be consistently estimated.

Second, like Caner (2005) (see also Hahn and Kuersteiner (2002) for the special case of linear 2SLS), we focus on the case dubbed *nearly-weak identification*, where the drifting DGP introduces a limit rank deficiency reached at a rate slower than root-T: this allows consistent estimation of all parameters, but at rates possibly slower than usual. It is then all the more important to identify the different directions in the parameter space endowed with the different rates. We consistently estimate these directions without assuming that the rates slower than root-T are known. We only maintain the assumption that the moment conditions responsible for approximate rank deficiency have been detected. Practically, this either may be thanks to prior economic knowledge (like market efficiency responsible for the weakness of instruments built from past returns in asset pricing models) or suggested by a preliminary study of the lack of steepness of the GMM objective function around plausible values of the structural parameters. Note that we only consider asymptotic rank deficiency such that all the rates of convergence of GMM estimators, possibly slower than root-T, are at least larger than $T^{1/4}$. The first order under-identification case of Sargan (1983), producing GMM estimators converging at rates $T^{1/4}$, can then be seen as a limit case of our approach. This is in sharp contrast with the weak instrument case *à la Stock and Wright* (2000) where the asymptotic rank deficiency is reached at a rate as fast as root-T: GMM estimators are not even consistent. The fact that all the GMM estimators are consistent with well-defined rates of convergence, albeit possibly unknown and slower than root-T, allows us to validate standard asymptotic testing approaches like Wald test or GMM-LM test of Newey and West (1987). In contrast with Kleibergen (2005), we do not need to modify the standard formulas for the LM test. Moreover, our approach is more general than Kleibergen (2005) since we explicitly take into account the possible simultaneous occurrence, in a given set of moment conditions, of heterogeneous rates of convergence.

As far as technical tools for asymptotic theory are concerned, we borrow to three recent developments in econometric theory.

First, as stressed by Stock and Wright (2000), (nearly)-weak identification in nonlinear settings makes asymptotic theory more involved than in the linear case because the occurrence of unknown parameters and observations in the moment conditions are not additively separable. Lee's (2004) minimum distance estimation with heterogeneous rates of convergence, albeit nonlinear, is also kept simple by this kind of additive separability. By contrast, this non-separability makes, in general, necessary resorting to a functional central limit theorem applied to the GMM objective function viewed as an empirical process indexed by unknown parameters.

Second, our approach to Wald testing with heterogeneous rates of convergence must be related to the former contribution of Lee (2005). The key issue is the following: when several directions (to be tested) in the parameter space are estimated at slow rates, while some linear combinations of them may be estimated at faster rates, a perverse asymptotic singularity is introduced and invalidates the common delta theorem. This situation, rather similar in spirit to cointegration, leads Lee (2005) to maintain an additional assumption for Wald testing. We are able to relax Lee's (2005) condition and to confirm that the common Wald test methodology always work, albeit with possibly nonstandard rates of convergence against sequences of local alternatives. The trick is again to consider a convenient rotation in the parameter space. Note that this issue makes even more important our extension of Kleibergen's (2005) setting to allow for different rates of convergence simultaneously.

A third debt to acknowledge is with respect to Andrews (1994, 1995) MINPIN estimators[1] and to Gagliardini, Gouriéroux and Renault (2005) XMM (Extended Method of Moments) estimators as well. Like them, we observe that GMM-like asymptotic variance formulas remain valid for strongly identified directions when slowly identified directions are estimated at rates faster than $T^{1/4}$. Rates even slower than that would imply a perverse contamination of the estimators of the standard directions by poorly identified nuisance parameters. In this respect, our approach should rather be dubbed *nearly-strong identification*. Of course, by doing so, we may renounce to capture severe weak identification cases arising even when the sample size is very large (see e.g Angrist and Krueger (1991)). However, our approach provides the empirical economist with estimation and inference procedures that are valid with the standard formulas, while warning her about rates of convergence in some specific directions that may be slower than the standard root-T. Moreover, these directions (strong and

[1] MINPIN estimators are defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator.

weak) can be disentangled and consistently estimated without modifying the overall rates of convergence of the implied linear combinations of structural parameters.

The chapter is organized as follows. Section 2 precisely defines our nearly-weak identification setting and proves consistency of both point GMM estimators of structural parameters θ and set estimators, that are equivalent to LM-tests of null hypotheses $\theta = \theta^0$. With nearly-weak global identification, consistency of point estimation rests upon an empirical process approach for the moment conditions, whereas set estimation rests upon nearly-weak local identification, characterized in terms of the expected Jacobian of the moment conditions. Our integrated framework restores the coherency between the two possible points of view about weak identification, global and local. In section 3, we show how to disentangle and to estimate the directions with different rates of convergence. We also prove the asymptotic normality of well-suited linear combinations of the structural parameters. The issue of Wald testing is addressed in section 4 while section 5 explicitly relates our setting to examples of weak identification well-studied in the literature. Section 6 is devoted to a couple of Monte-Carlo illustrations for two toys models: single-equation linear IV model and CCAPM.
All the proofs and figures are gathered in the appendix[2].

2 Consistent point and set GMM estimators

This section shows that a standard GMM approach works both for consistent point and set estimation, the latter through a score type test statistic. Typically, all the components of the parameters of interest are simultaneously estimated and tested without *a priori* knowledge of their heterogenous patters of identification.

2.1 Nearly-weak global identification

Let θ be a p-dimensional parameter vector with true (unknown) value θ^0, assumed in the interior of the compact parameter space Θ. The true parameter value satisfies the K equations,

$$E\left[\omega_t(\theta^0)\right] = 0 \tag{2.1}$$

[2]Most of the theoretical results are obtained in a more general context in a technical companion paper Antoine and Renault (2007).

with $\phi(.)$ some known functions. We have at our disposal a sample of size T, and we can calculate $\phi_t(\theta)$ for any value of the parameter in Θ and for every $t = 1, \cdots, T$.

Standard GMM estimation defines its estimator $\hat{\theta}_T$ as follows:

Definition 2.1. *Let Ω_T be a sequence of symmetric positive definite random matrices of size K which converges in probability towards a positive definite matrix Ω. A GMM estimator $\hat{\theta}_T$ of θ^0 is then defined as:*

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} Q_T(\theta) \quad \textit{where} \quad Q_T(\theta) \equiv \bar{\phi}_T'(\theta)\Omega_T\bar{\phi}_T(\theta) \tag{2.2}$$

with $\bar{\phi}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\phi_t(\theta)$, the empirical mean of the moment restrictions.

Standard GMM asymptotic theory assumes that, for $\theta \neq \theta^0$, $\bar{\phi}_T(\theta)$ converges in probability towards its nonzero expected value because of some uniform law of large numbers. We consider here a more general situation where $\bar{\phi}_T(\theta)$ may converge towards zero even for $\theta \neq \theta^0$. And we show how this can be interpreted as identification issues.

More precisely, we imagine that we have here two groups of moment restrictions: one standard for which the empirical counterpart converges at the standard (usual) rate of convergence $\sqrt{T}$ and a weaker one for which the empirical counterpart still converges but potentially at a slower rate λ_T. At this stage, it is essential to stress that identification is going to be maintained (but through higher order asymptotic developments). More formally, we have k_1 standard moment restrictions such that

$$\sqrt{T}\left[\bar{\phi}_{1T}(\theta) - \rho_1(\theta)\right] = \mathcal{O}_P(1) \tag{2.3}$$

and k_2 $(= K - k_1)$ weaker moment restrictions such that

$$\sqrt{T}\left[\bar{\phi}_{2T}(\theta) - \frac{\lambda_T}{\sqrt{T}}\rho_2(\theta)\right] = \mathcal{O}_P(1) \text{ where } \lambda_T = o(\sqrt{T}) \text{ and } \lambda_T \xrightarrow{T} \infty \tag{2.4}$$

with $[\rho_1'(\theta)\ \rho_2'(\theta)] = 0 \Leftrightarrow \theta = \theta^0$.

λ_T measures the degree of weakness of the second group of moment restrictions. The corresponding component $\rho_2(.)$ is squeezed to zero and $Plim\left[\bar{\phi}_{2T}(\theta)\right] = 0$ for all $\theta \in \Theta$. Thus, the probability limit of $\bar{\phi}_T(\theta)$ does not allow to discriminate between θ^0 and any $\theta \in \Theta$. In

such a context, identification is a combined property of the functions $\phi_t(\theta)$ and $\rho(\theta)$ and the asymptotic behavior of λ_T. The maintained identification assumption is the following:

Assumption 1. *(Identification)*
(i) $\rho(.)$ is a continuous function from a compact parameter space $\Theta \subset \mathbb{R}^p$ into $\mathbb{R}^K$ such that

$$\rho(\theta) = 0 \iff \theta = \theta^0$$

(ii) The empirical process $(\Psi_T(\theta))_{\theta\in\Theta}$ obeys a functional central limit theorem:

$$\Psi_T(\theta) \equiv \sqrt{T}\begin{bmatrix} \overline{\phi}_{1T}(\theta) - \rho_1(\theta) \\ \overline{\phi}_{2T}(\theta) - \frac{\lambda_T(\theta)}{\sqrt{T}}\rho_2(\theta) \end{bmatrix} \Rightarrow \Psi(\theta)$$

where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero.
(iii) λ_T is a deterministic sequence of positive real numbers such that

$$\lim_{T\to\infty} \lambda_T = \infty \text{ and } \lim_{T\to\infty} \frac{\lambda_T}{\sqrt{T}} = 0$$

Following Stock and Wright (2000), assumption 1 reinforces the standard central limit theorem written for moment conditions at the true value ($\theta = \theta^0$) by maintaining a functional central limit theorem on the whole parameter set Θ. Stock and Wright (2000) use this framework to address the weak identification case corresponding to $\lambda_T = 1$. By contrast, as Hahn and Kuersteiner (2002) and Caner (2005), we focus here on nearly-weak identification where λ_T goes to infinity albeit slower than $\sqrt{T}$. Note that the standard strong identification case corresponds to $\lambda_T = \sqrt{T}$. The above functional central limit theorem[3] allows us to get a consistent GMM estimator, even in case of nearly-weak identification[4].

Theorem 2.1. *(Consistency of $\hat{\theta}_T$)*
Under assumption 1, any GMM estimator $\hat{\theta}_T$ like (2.2) is weakly consistent.

[3]Note that the asymptotic normality assumption is not necessary at this stage. In general, it might be replaced by some tightness assumption on $\Psi_T(.)$. See Antoine and Renault (2007).

[4]As stressed by Stock and Wright (2000) the uniformity in θ provided by the functional central limit theorem is crucial in case of nonlinear nonseparable moment conditions, that is when the occurrences of θ and the observations in the moment conditions are not additively separable. By contrast, Hahn and Kuersteiner (2002) (linear case) and Lee (2004) (separable case) do not need to resort to a functional central limit theorem.

Besides the fact that all the components of the parameter of interest θ are consistently estimated, it is worth stressing another difference with Stock and Wright (2000). We do not assume the *a priori* knowledge of a partition $\theta = (\alpha' \ \beta')'$, where α is strongly identified and β (nearly)-weakly identified. By contrast, nearly-weak identification is produced by the rates of convergence of the moment conditions. More precisely, assumption 1 implies that, for the first set of moment conditions, we have (as for standard GMM),

$$\rho_1(\theta) = \underset{T\to\infty}{Plim}\ \bar{\phi}_{1T}(\theta)$$

whereas we only have for the second set of moment conditions

$$\rho_2(\theta) = \underset{T\to\infty}{Plim}\ \frac{\sqrt{T}}{\lambda_T}\bar{\phi}_{2T}(\theta)$$

It will be shown that this framework nests Stock and Wright (2000), Hahn and Kuersteiner (2002) and Caner (2005). More precisely, a rotation in the parameter space will allow us to identify some strongly identified directions and some others, only (nearly)-weakly identified. Subsection 2.2 below shows that the above rates of convergence naturally induce rates of convergence for the Jacobian matrices. This enables us to encompass the framework of Kleibergen (2005).

2.2 Nearly-weak local identification

As already explained, we simultaneously consider two rates of convergence to characterize the asymptotic behavior of the sample moments $\bar{\phi}_T(\theta)$ and the induced singularity issues in the sample counterparts of the estimating functions $\rho(\theta)$. In this respect, we differ from Sargan (1983) since we maintain the first-order identification assumption:

Assumption 2. *(First-order identification)*
(i) $\rho(.)$ *is continuously differentiable on the interior of* Θ *denoted as* $int(\Theta)$.
(ii) $\theta^0 \in int(\Theta)$.
(iii) The $(K \times p)$*-matrix* $[\partial\rho(\theta)/\partial\theta']$ *has full column rank* p *for all* $\theta \in \Theta$.
(iv) $\left[\begin{array}{cc} \frac{\partial\rho_1'(\theta^0)}{\partial\theta} & \frac{\partial\rho_2'(\theta^0)}{\partial\theta} \end{array}\right] = \underset{T\to\infty}{Plim}\left[\begin{array}{cc} \frac{\partial\bar{\phi}_{1T}'(\theta^0)}{\partial\theta} & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\bar{\phi}_{2T}'(\theta^0)}{\partial\theta} \end{array}\right]$
(v) $\sqrt{T}\left[\frac{\partial\bar{\phi}_{1T}(\theta^0)}{\partial\theta'} - \frac{\partial\rho_1(\theta^0)}{\partial\theta'}\right] = \mathcal{O}_P(1)$

The identification issue is not raised by rank deficiency of the moment conditions but by the rates of convergence. In other words, the implicit assumption in Kleibergen (2005) (see the proof of his theorem 1 page 1122) that Jacobian matrices may have non-standard rates of convergence is made explicit in our framework. Assumptions 2(iv) and (v) are the natural extensions of assumption 1 on Jacobian matrices. Typically, Kleibergen (2005) maintains assumption 2(v) through a joint asymptotic normality assumption on $\overline{\phi}_T(\theta^0)$ and $\left[\partial\overline{\phi}_T(\theta^0)/\partial\theta'\right]$ (see his assumption 1).

While global identification (assumption 1) provides a consistent estimator of θ, local identification (assumption 2) provides an asymptotically consistent confidence set at level $(1-\alpha)$ or, equivalently, an asymptotically consistent test at level α for any simple hypothesis $H_0 : \theta = \theta_0$[5]. A score test approach, as defined in Newey and West (1987), does not resort to the asymptotic distributions of the estimators:

Theorem 2.2. *(Score test)*
The score statistic for testing $H_0 : \theta = \theta_0$ is defined as

$$LM_T(\theta_0) = \frac{T}{4}\frac{\partial Q_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial Q'_T(\theta_0)}{\partial\theta}$$

where S_T is a standard consistent estimator of the long-term covariance matrix[6].
Under H_0 and assumptions 1 and 2, $LM_T(\theta_0)$ has a $\chi^2(p)$ limit distribution.

In sharp contrast with Kleibergen (2005), we do not need to modify the standard score test statistic to replace the Jacobian of the moment conditions by their projections on the orthogonal space of the moment conditions. The reason for this maintained simplicity is that, in our nearly-weakly identified case,

$$\frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'}$$

has a deterministic limit which does not introduce any perverse correlations. By contrast, in the weakly identified case considered by Kleibergen (2005) (or $\lambda_T = 1$), the relevant limit of the sequence of Jacobian matrices is Gaussian. In this latter case, the limiting behavior

[5]Note that in general θ_0 might be different from the true (unknown) value of the parameter θ^0.

[6]Note that a consistent estimator S_T of the long-term covariance matrix $S(\theta_0)$ of $\Psi(\theta_0)$ can be built in the standard way (see in general Hall (2005)) from a preliminary inefficient GMM estimator $\tilde{\theta}_T$ of θ. However, under the null, one may simply choose $\tilde{\theta}_T = \theta_0$.

of $\left[\partial\bar{\phi}_T(\theta^0)/\partial\theta'\right]$ is not independent of the limiting behavior of $\left[\bar{\phi}_T(\theta^0)\right]$ so the limiting distribution of the GMM score test statistic depends on nuisance parameters (see Stock and Wright (2000)). Of course, the advantage of the K-statistic proposed by Kleibergen (2005) is to be robust in the limit case $\lambda_T = 1$ while, for us, λ_T must always converge towards infinity albeit possibly very slowly.

It is essential to realize that although the standard score test statistic has the common $\chi^2(p)$ distribution under the null, it works rather differently. Basically,

$$\left[\frac{\partial\bar{\phi}_T'(\theta^0)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\theta^0)}{\partial\theta'}\right] \tag{2.5}$$

is an asymptotically singular matrix since

$$\underset{T\to\infty}{Plim}\left[\frac{\partial\bar{\phi}_{2T}(\theta^0)}{\partial\theta'}\right] = \lim_{T\to\infty}\left[\frac{\lambda_T}{\sqrt{T}}\frac{\partial\rho_2(\theta^0)}{\partial\theta'}\right] = 0$$

The proof of theorem 2.2 shows that the standard formula is actually recovered by well-suited matricial scalings of $\left[\partial Q_T(\theta^0)/\partial\theta'\right]$ and (2.5). The ultimate cancelation of these scalings must not conceal that testing parameter in GMM without assuming they are strongly identified requires a specific theory. It is in particular important to realize that both strong and (nearly)-weak identification may show up together in a given set of moment conditions. Note that this is immaterial as far as practical formulas for score testing are concerned. However, we show below that it has a dramatic impact on the power against local alternatives[7].

Another difference with Kleibergen (2005) is that our score test is consistent in all directions. Actually, ignoring the limit case ($\lambda_T = 1$) of weak identification allows us to write down consistent confidence sets and score tests. In terms of local alternatives, we get consistency at least at rate λ_T thanks to the following result:

Theorem 2.3. *(Rate of convergence)*
Under assumptions 1 and 2(i) to (iii), we have:

$$\left\|\hat{\theta}_T - \theta^0\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

[7] Kleibergen (2005) considers a simpler setting since he does not allow for two different kinds of identification (strong and weak) to be considered simultaneously (see the proof of his theorem 1). In addition, a full rank condition seems to be implicitly maintained in Kleibergen's proof.

In the remaining of the essay, we precisely focus on the identification of directions of local alternatives where consistency is kept at the standard rate $\sqrt{T}$.

3 Rates of convergence and asymptotic normality

In this section, we start with a kind of *rotation* in the parameter space which allows us to disentangle the rates of convergence. More precisely, some special linear combinations of θ are actually estimated at the standard rate of convergence $\sqrt{T}$, while some others are still estimated at the slower rate λ_T. This is formalized by a central limit theorem which allows the practitioner to apply the common GMM formula without knowing *a priori* the identification pattern.

3.1 Separation of the rates of convergence

We face the following situation:

(i) Only k_1 equations (defined by $\rho_1(.)$) have a sample counterpart which converges at the standard rate $\sqrt{T}$. These can be used in a standard way. Unfortunately, we have in general a reduced rank problem: $[\partial \rho_1(\theta^0)/\partial \theta']$ is not full column rank. Its rank s_1 is strictly smaller than p and the first set of equations cannot identify θ. Intuitively, it can only identify s_1 directions in the p-dimensional space of parameters.

(ii) The k_2 remaining equations (defined by $\rho_2(.)$) should be used to identify the remaining $s_2(=p-s_1)$ directions[8]. However this additional identification comes at the slower rate λ_T.

We already have the intuition that the parameter space is going to be separated into two subspaces: the first one (defined through $\rho_1(.)$) collects s_1 standard directions and the second one (defined through $\rho_2(.)$) gathers the remaining (slow) directions. We now make this separation much more precise by defining a reparametrization. Each of the above subspaces is actually characterized as the range of a full column rank matrix: respectively the (p, s_1)-matrix R_1^0 and the $(p, p-s_1)$-matrix R_2^0.

[8]Recall that, by assumption, our set of moment conditions enables the identification of the entire vector of parameters θ.

Since R_2^0 characterizes the set of slow directions, it is natural to define it via the null space of $[\partial \rho_1'(\theta^0)/\partial\theta]$, or, in other words, everything that is not identified in a standard way (through $\rho_1(.)$):

$$\frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_2^0 = 0 \tag{3.1}$$

Then these $(p - s_1)$ (slow) directions are completed with the definition of the remaining s_1 directions as follows:

$$R^0 = \left[R_1^0\ R_2^0\right] \quad \text{and} \quad Rank\left[R^0\right] = p$$

Then R^0 is a nonsingular (p, p)-matrix that can be used as a matrix of a change of basis in $\mathbb{R}^p$. More precisely, we define the new parameter as $\eta = [R^0]^{-1}\theta$, that is

$$\theta = [R_1^0\ R_2^0] \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \begin{matrix} \updownarrow s_1 \\ \updownarrow p - s_1 \end{matrix}$$

We will see in the next subsection that this reparametrization precisely isolates the two rates of convergence by defining two subsets of directions, each of them associated with a rate of convergence. The reparametrization also shows that, in general, there is no hope to get standard asymptotic normality of some components of the estimator $\hat{\theta}_T$ of θ^0. The reason is simple: after a standard expansion of the first-order conditions, $\hat{\theta}_T$ now appears as asymptotically equivalent to some linear transformations of $\overline{\phi}_T(\theta)$ which are likely to mix up the two rates. In other words, all components of $\hat{\theta}_T$ might be contaminated by the slow rate of convergence. Hence the main advantage of the reparametrization is precisely to separate these two rates. In section 5.1 where we carefully compare our theory with Stock and Wright (2000), we provide conditions under which some components of $\hat{\theta}_T$ are (by chance) converging at the standard rate. And this is exactly what is assumed *a priori* by Stock and Wright (2000) when they separate the structural parameters into one *standard-converging* group and one *slower-converging* one.

The reparametrization may not be feasible in practice since the matrix R^0 depends on the true unknown value of the parameter θ^0. However, we can still deduce a feasible inference strategy.

3.2 Efficient estimation

To be able to get an asymptotic normality result on the new set of parameters, we need some technical assumptions and preliminary results. More details can be found in the technical companion paper by Antoine and Renault (2007).

It is worth noting that, albeit with a mixture of different rates, the Jacobian matrix of moment conditions has a consistent sample counterpart. Let us first define the following (p, p) block diagonal scaling matrix $\tilde{\Lambda}_T$, where Id_r denotes the identity matrix of size r:

$$\tilde{\Lambda}_T = \begin{pmatrix} \sqrt{T} Id_{s_1} & 0 \\ 0 & \lambda_T Id_{s_2} \end{pmatrix}$$

As it can be seen in the proof of theorem 2.2, assumption 2 ensures that:

$$\sqrt{T} \frac{\partial \bar{\phi}_T(\theta^0)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \xrightarrow{P} J^0 \quad \text{with} \quad J^0 \equiv \frac{\partial \rho(\theta^0)}{\partial \theta'} R^0 \tag{3.2}$$

where J^0 is the (K, p) block diagonal matrix with its two blocks respectively defined as the (k_i, s_i) matrices $[\partial \rho_i(\theta^0)/\partial \theta' \; R_i^0]$ for $i = 1, 2$. Note that the coexistence of two rates of convergence (λ_T and $\sqrt{T}$) implies zero north-east and south-west blocks for J^0.

Moreover to derive the asymptotic distribution of the GMM estimator $\hat{\theta}_T$ (through well-suited Taylor expansions of the first order conditions), the above convergence towards J^0 needs to be fulfilled even when the true value θ^0 is replaced by some preliminary consistent estimator θ_T^*. Hence, Taylor expansions must be robust to a λ_T-consistent estimator, the only rate guaranteed by theorem 2.3. This situation is rather similar to the one studied in Andrews (1994) for the so-called MINPIN estimator[9]. We do not want the slow convergence of some directions to contaminate the standard convergence of the others (see theorem 3.1 below): more precisely, we need to ensure that the slow rate λ_T does not modify the relative orders of magnitude of the different terms of the Taylor expansions. As Andrews (1995 p563) does for nonparametric estimators, we basically need to assume that our nearly-weakly identified

[9] MINPIN estimators are estimators defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator. These nuisance parameters are estimated at slower rates and one wants to prevent their distributions to contaminate the asymptotic distribution of the parameters of interest.

directions are estimated at a rate faster than $(T^{1/4})$.[10] In addition, we want as usual uniform convergence of sample Hessian matrices. This leads us to maintain the following assumption:

Assumption 3. *(Taylor expansions)*
(i) $\lim_{T\to\infty}\left[\frac{\lambda_T^2}{\sqrt{T}}\right]=\infty$
(ii) $\bar{\phi}_T(\theta)$ *is twice continuously differentiable on the interior of* Θ *and is such that:*

$$\forall\, 1\leq k\leq k_1\ \frac{\partial^2\bar{\phi}_{1T,k}(\theta)}{\partial\theta\partial\theta'}\xrightarrow{P}H_{1,k}(\theta)\quad \textit{and}\quad \forall\, 1\leq k\leq k_2\ \frac{\sqrt{T}}{\lambda_T}\frac{\partial^2\bar{\phi}_{2T,k}(\theta)}{\partial\theta\partial\theta'}\xrightarrow{P}H_{2,k}(\theta)$$

uniformly on θ *in some neighborhood of* θ^0*, for some* (p,p) *matricial function* $H_{i,k}(\theta)$ *for* $i=1,2$ *and* $1\leq k\leq k_i$.

While common weak identification corresponds to $\lambda_T=1$ and strong identification to $\lambda_T=\sqrt{T}$, our approach in the rest of the essay is actually a rather *nearly-strong* one since we assume λ_T strictly between $T^{1/4}$ and $\sqrt{T}$.[11]

Up to unusual rates of convergence, we get a standard asymptotic normality result for the new parameter η:

Theorem 3.1. *(Asymptotic Normality)*
(i) Under assumptions 1 to 3, the GMM estimator $\hat{\theta}_T$ *defined by (2.2) is such that:*

$$\tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T-\theta^0\right)\xrightarrow{d}\mathcal{N}\left(0,\left[J^{0\prime}\Omega J^0\right]^{-1}J^{0\prime}\Omega S(\theta^0)\Omega J^0\left[J^{0\prime}\Omega J^0\right]^{-1}\right)$$

(ii) Under assumptions 1 to 3, the asymptotic variance displayed in (i) is minimal[12] *when the GMM estimator* $\hat{\theta}_T$ *is defined with a weighting matrix* Ω_T *being a consistent estimator of* $\Omega=[S(\theta^0)]^{-1}$*:*

$$\tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T-\theta^0\right)\xrightarrow{d}\mathcal{N}\left(0,\left[J^{0\prime}[S(\theta^0)]^{-1}J^0\right]^{-1}\right)$$

[10] More details on the link between Andrews (1994, 1995) and this setting might also be found in Antoine and Renault (2007).

[11] It is worth reminding that the score test derived in section 2 is valid for λ_T arbitrarily close to the weak identification case.

[12] Note that efficiency is implicitly considered here for the given set of moment restrictions $\bar{\phi}_T(.)$. In section 3.3, we study the consequences of rewriting the moment conditions.

Note that $\hat{\eta}_T = [R^0]^{-1}\hat{\theta}_T$ can be interpreted as a consistent estimator of $\eta^0 = [R^0]^{-1}\theta^0$. Of course it is not feasible since R^0 is unknown. The issue of plugging in a consistent estimator of R^0 will be addressed in section 3.4. For the moment, our focus of interest are the implied rates of convergence for inference about θ. Since

$$\hat{\theta}_T = R_1^0\hat{\eta}_{1,T} + R_2^0\hat{\eta}_{2,T}$$

a linear combination $a'\hat{\theta}_T$ of the estimated parameters of interest will be endowed with a $\sqrt{T}$ rate of convergence of $\hat{\eta}_{1,T}$ if and only if $a'R_2^0 = 0$, that is a belongs to the orthogonal space of the range of R_2^0. By virtue of equation (3.1) the latter property means that a is spanned by the columns of the matrix $[\partial\rho_1'(\theta^0)/\partial\theta]$. In other words, $a'\theta$ is strongly identified if and only if it is identified by the first set of moment conditions $\rho_1(\theta) = 0$.

As far as inference about θ is concerned, several practical implications of theorem 3.1 are worth mentioning. Up to the unknown matrix R^0 and the unknown rate of convergence λ_T (which appears in $\tilde{\Lambda}_T$), a consistent estimator of the asymptotic covariance matrix $\left(J^{0\prime}\left[S(\theta^0)\right]^{-1}J^0\right)^{-1}$ is[13]

$$T^{-1}\tilde{\Lambda}_T[R^0]^{-1}\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}[R^{0\prime}]^{-1}\tilde{\Lambda}_T \tag{3.3}$$

where S_T is a standard consistent estimator of the long-term covariance matrix[14]. From theorem 3.1, for large T, $\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)$ behaves like a gaussian random variable with mean zero and variance (3.3). One may be tempted to deduce that $\sqrt{T}(\hat{\theta}_T - \theta^0)$ behaves like a gaussian random variable with mean 0 and variance

$$\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \tag{3.4}$$

And this would give the feeling that we are back to standard GMM formulas of Hansen (1982). As far as practical purposes are concerned, this intuition is correct: note in particular that the knowledge of R^0 is not necessary to perform inference. However, from a theoretical point of view, this is a bit misleading. First since in general all components of $\hat{\theta}_T$ converge

[13]This directly follows from lemma B in the appendix.

[14]Note that a consistent estimator of S_T of the long-term covariance matrix $S(\theta^0)$ can be built in the standard way (see in general Hall(2005)) from a preliminary inefficient GMM estimator $\bar{\theta}_T$ of θ.

asymptotic variance (3.4) is akin to refer to the inverse of an asymptotically singular matrix. Second, for the same reason, (3.4) is not an estimator of the standard population matrix

$$\left[\frac{\partial\rho'(\theta^0)}{\partial\theta}\left[S(\theta^0)\right]^{-1}\frac{\partial\rho(\theta^0)}{\partial\theta'}\right]^{-1} \tag{3.5}$$

To conclude, if inference about θ is technically more involved than one may believe at first sight, it is actually very similar to standard GMM formulas from a pure practical point of view. In other words, if a practitioner is not aware of the specific framework with moment conditions associated with several rates of convergence (coming, say, from the use of instruments of different qualities) then she can still provide reliable inference by using standard GMM formulas. In this respect, we generalize Kleibergen's (2005) result that inference can be performed without *a priori* knowledge of the identification setting. However as already mentioned in section 2, we are more general that Kleibergen (2005) since we allow moment conditions to display simultaneously different identification patterns[15].

Finally, the standard score test defined in theorem 2.2 may be completed by a classical overidentification test:

Theorem 3.2. *(J-test)*
Under assumptions 1 to 3, if Ω_T is a consistent estimator of $\left[S(\theta^0)\right]^{-1}$, then

$$TQ_T(\theta_T) \xrightarrow{d} \chi^2_{K-p}$$

3.3 Orthogonalization of the moment restrictions

In this section, we investigate the consequences of transforming the moment restrictions to estimate the standard and slow directions. Since we deal simultaneously with standard and weaker moment conditions, we cannot consider any linear combination of the restrictions. In particular, we can only consider transformations preserving the central limit theorem in Assumption 1, and the fragile information of the weaker moment restrictions. Any valid

[15] For sake of notational simplicity, we only consider in this essay one speed of nearly-weak identification λ_T. The reader interested in working with an arbitrary number of different speeds might use the general framework of Antoine and Renault (2007).

transformation of the moment conditions, or transformation that does not affect the true moment conditions $\rho_1(.)$ and $\rho_2(.)$, can be written as follows:

$$\begin{bmatrix} \overline{\phi}_{1T}^{H}(\theta^0) \\ \overline{\phi}_{2T}^{H}(\theta^0) \end{bmatrix} = \begin{bmatrix} \overline{\phi}_{1T}(\theta^0) + H\overline{\phi}_{2T}(\theta^0) \\ \overline{\phi}_{2T}(\theta^0) \end{bmatrix} \tag{3.6}$$

for some (k_1, k_2)- matrix H that may depend on the sample size T and the true unknown parameter θ^0.

A linear transformation of interest in the literature is the orthogonalization: the standard moment conditions are replaced by the residuals of their regression on the set of weaker moment conditions. The set of the empirical mean of the moment conditions $[\overline{\phi}_{1T}' \ \overline{\phi}_{2T}']'$ is replaced by $[\tilde{\phi}_{1T}' \ \tilde{\phi}_{2T}']'$ defined as follows:

$$\begin{bmatrix} \overline{\phi}_{1T}(\theta^0) - Cov\left(\sqrt{T}\overline{\phi}_{1T}(\theta^0), \sqrt{T}\overline{\phi}_{2T}(\theta^0)\right)\left[Var\left(\sqrt{T}\overline{\phi}_{2T}(\theta^0)\right)\right]^{-1}\overline{\phi}_{2T}(\theta^0) \\ \overline{\phi}_{2T}(\theta^0) \end{bmatrix} \tag{3.7}$$

Note that we still have the same central limit theorem, only the asymptotic variance is modified:

$$\sqrt{T}\begin{bmatrix} \tilde{\phi}_{1T}(\theta^0) - \rho_1(\theta^0) \\ \tilde{\phi}_{2T}(\theta^0) - \frac{\lambda_T}{\sqrt{T}}\rho_2(\theta^0) \end{bmatrix} \Rightarrow \tilde{\Psi}(\theta^0)$$

where $\tilde{\Psi}(\theta)$ is a Gaussian random variable with mean zero and block diagonal matrix Σ^0 with respective blocks $\Sigma_1^0 = S_1^0 - S_{12}^0[S_2^0]^{-1}S_{21}^0$ and $\Sigma_2^0 = S_2^0$.

The following theorem compares the asymptotic variances of the estimators associated to the original set of moment conditions $\hat{\eta}_T$ and to the orthogonalized one denoted as $\tilde{\eta}_T$:

Theorem 3.3. *(Orthogonalization)*
Consider the new parameter $\eta = [R^0]^{-1}\theta$. The two estimators obtained respectively from the GMM estimator associated with the original set of moment conditions $\hat{\theta}_T$ and from the GMM estimator associated with the orthogonalized set of moment conditions (3.7) $\tilde{\theta}_T$ are such that:
i) The orthogonalization improves the estimation of the standard directions (in terms of asymptotic variance matrix) ie

$$AVar\left[\hat{\eta}_{1T}\right] \geq\geq AVar\left[\tilde{\eta}_{1T}\right]$$

ii) The orthogonalization deteriorates the estimation of the slow directions (in terms of asymptotic variance matrix) ie

$$AVar\left[\hat{\eta}_{2T}\right] \leq\leq AVar\left[\tilde{\eta}_{2T}\right]$$

where $\leq\leq$ and $\geq\geq$ denote the comparisons between matrixes.

We show that the orthogalization of any valid set of moment restrictions (3.6) leads to the same set (3.7):

Proposition 3.4. *Any set of valid moment conditions like (3.6) leads to the same orthogonalized set of moment conditions (3.7).*

Denote by η_T^H the (transformed) estimator associated to the above moment conditions (3.6). We now show that among all the valid transformations, the orthogonalization is the best for the standard directions and the worse the slow ones.

Corollary 3.5. *Consider the new parameter $\eta = [R^0]^{-1}\theta$. The two estimators obtained respectively from the GMM estimator associated with the transformed set of moment conditions (3.6) $\tilde{\theta}_T^H$ and from the GMM estimator associated with the orthogonalized set of moment conditions (3.7) $\tilde{\theta}_T$ are such that:*
i) The orthogonalization is the best (valid) transformation of the moment conditions in terms of the efficiency of the standard directions ie

$$AVar\left[\eta_{1T}^H\right] \geq\geq AVar\left[\tilde{\eta}_{1T}\right]$$

ii) The orthogonalization is the worst (valid) transformation of the moment conditions in terms of the efficiency of the slow directions ie

$$AVar\left[\eta_{2T}^H\right] \leq\leq AVar\left[\tilde{\eta}_{2T}\right]$$

3.4 Estimating the strongly-identified directions

In this subsection, we provide a feasible way to estimate the strongly-identified directions in the parameter space. Recall that these directions have been identified through the following reparametrization,

$$[R^0]^{-1}\theta \equiv \begin{pmatrix} A^0\theta \\ B^0\theta \end{pmatrix}$$

where $(A^0\theta)$ represent the s_1 standard directions while $(B^0\theta)$ are the weaker ones. In general, this reparametrization is unfeasible since it depends on the unknown value of the parameter θ^0. To make this approach feasible, the key lemma which allows us to replace the above directions by their estimated counterparts is the following:

Lemma 3.6. *(Estimating the rotation in the parameter space)*
Under assumptions 1 to 3, if the vector

$$\begin{bmatrix} \sqrt{T}A\left(\hat{\theta}_T - \theta^0\right) \\ \lambda_T B\left(\hat{\theta}_T - \theta^0\right) \end{bmatrix}$$

is asymptotically gaussian and if $\hat{A}$ and $\hat{B}$ are consistent estimators of A and B such that

$$\left\|\hat{A} - A\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right) \quad \textit{and} \quad \left\|\hat{B} - B\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

then the vector

$$\begin{bmatrix} \sqrt{T}\hat{A}\left(\hat{\theta}_T - \theta^0\right) \\ \lambda_T \hat{B}\left(\hat{\theta}_T - \theta^0\right) \end{bmatrix}$$

is asymptotically gaussian.

In the proof of lemma 3.6, our *nearly-strongly* point of view is the essential key to keep the $\sqrt{T}$ convergence while relevant directions are only estimated at the slower rate λ_T: that is λ_T is small in front of $\sqrt{T}$ but large in front of $\left[T^{1/4}\right]$.[16]

4 Wald testing

In this section, we focus on testing a system of q restrictions about θ, say the null hypothesis $H_0 : g(\theta) = 0$, where $g(.)$ is a function from Θ to $\mathbb{R}^q$ continuously differentiable on the interior of Θ.

First, working under the null may conduct us to dramatically revisit the reparametrization $\eta = \left[R^0\right]^{-1}\theta$ defined in section 3. Typically additional information may lead us to define

[16] As already mentioned, this is very similar in spirit to MINPIN estimators of Andrews (1994, 1995).

differently the linear combinations of θ estimated respectively with standard and slow rates of convergence. To circumvent this difficulty, we do not consider any constrained estimator and we focus on Wald testing. Caner (2005) overlooks this complication and derives the standard asymptotic equivalence results for the trinity of tests. This is because he only treats asymptotic testing when all the parameters converge at the same speed.

Second, as already explained, the main originality of this essay is to allow for the simultaneous treatment of different identification patterns. This more general point of view comes at a price when one wants to test. More precisely, we may face singularity issues when some tested restrictions estimated at the slow rate λ_T can be linearly combined so as to be estimated at the standard rate $\sqrt{T}$. Lee (2005) puts forward some high level assumptions (see his assumptions (R) and (G)) to deal with the asymptotic singularity problem. We show that our setting allows us to perform a standard Wald test even without maintaining Lee's (2005) high-level assumptions.

From our discussion in sections 2 and 3, we can guess that a Wald test statistic for H_0 can actually be written with a standard formula:

$$\zeta_T^W = T g'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial \theta'} \left[\frac{\partial \bar{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} \right\}^{-1} g(\hat{\theta}_T)$$

Recall the standard rank assumption ensuring that the Wald test statistic is asymptotically chi-square with q degrees of freedom:

$$Rank \left[\frac{\partial g(\theta)}{\partial \theta'} \right] = q \tag{4.1}$$

for all θ in the interior of Θ, or at least in a neighborhood of θ^0. As well known, this condition is not really restrictive since it is akin to say that, at least locally, the q restrictions under test are linearly independent. Unfortunately, the existence of different rates of convergence may introduce (asymptotically) some perverse multicolinearity between the q estimated constraints.

The counterexample below points out the key issue.

Example 4.1. *(Counterexample)*
Assume that we want to test $H_0 : g(\theta) = 0$ *with* $g(\theta) = [g_j(\theta)]_{1 \leq j \leq q}$ *and none of the* q *vectors*

$\partial g_j(\theta^0)/\partial\theta$, $j = 1, \cdots, q$ belongs to $Im[\partial\rho_1'(\theta^0)/\partial\theta]$[17]. Then the extension of the standard argument for Wald test would be to say that, under the null, $\lambda_T g(\hat{\theta}_T)$ behaves asymptotically like $\partial g(\theta^0)/\partial\theta'\lambda_T(\hat{\theta}_T - \theta^0)$, that is for large T, $\lambda_T g(\hat{\theta}_T)$ behaves like a gaussian

$$\mathcal{N}\left(0, \frac{\partial g(\theta^0)}{\partial\theta'}\left[\frac{\partial\bar{\phi}_T'(\theta^0)}{\partial\theta}S(\theta^0)^{-1}\frac{\partial\bar{\phi}_T(\theta^0)}{\partial\theta'}\right]^{-1}\frac{\partial g'(\theta^0)}{\partial\theta}\right)$$

Imagine however that for some nonzero vector α,

$$\alpha'\frac{\partial g(\theta^0)}{\partial\theta'} = \sum_{j=1}^{q}\alpha_j\frac{\partial g_j(\theta^0)}{\partial\theta'}$$

belongs to $Im[\partial\rho_1'(\theta^0)/\partial\theta]$. Then (see comments after theorem 3.1) under the null $\sqrt{T}\alpha' g(\hat{\theta}_T)$ is asymptotically gaussian and thus

$$\lambda_T\alpha' g(\hat{\theta}_T) = \frac{\lambda_T}{\sqrt{T}}\sqrt{T}\alpha' g(\hat{\theta}_T) \xrightarrow{P} 0$$

In other words, even if the q constraints are locally linearly independent (ie $Rank[\partial g(\theta^0)/\partial\theta'] = q$) $\left[\lambda_T g(\hat{\theta}_T)\right]$ does not behave asymptotically like a gaussian with a non-singular variance matrix. This is the reason why deriving an asymptotically $\chi^2(q)$ distribution for the Wald test statistic is more involved than usual.

Lee (2005) avoids this kind of perverse asymptotic singularity by maintaining the following assumption:

Lee's (2005) assumption:
There exists a sequence of (q, q) invertible matrices D_T such that for any $\theta \in \Theta$

$$\underset{T\to\infty}{Plim}\left[D_T\frac{\partial g(\theta^0)}{\partial\theta'}R^0[\tilde{\Lambda}_T]^{-1}\right] = B_0$$

where B_0 is a (q, p) deterministic finite matrix of full row rank.

Lee's (2005) assumption clearly implies the standard rank condition (4.1). However, the converse is not true as it can be shown from the counterexample above[18]. And this is actually

[17]For any $(n \times m)$-matrix, $Im[M]$ represents the subspace of $\mathbb{R}^n$ generated by the column vectors of M. It is also referred to as $Col[M]$ and $Range[M]$.

[18]By contrast, in the case of only $q = 1$ constraint, Lee's assumption is trivially fulfilled.

above assumption implies that, under the null, $D_T g(\tilde{\theta}_T)$ behaves like $D_T \partial g(\theta^0)/\partial\theta'(\tilde{\theta}_T - \theta^0)$ that is like $B^0 \tilde{\Lambda}_T [R^0]^{-1}(\tilde{\theta}_T - \theta^0)$. From theorem 3.1, we know that $\tilde{\Lambda}_T [R^0]^{-1}(\tilde{\theta}_T - \theta^0)$ nicely behaves as an asymptotic gaussian distribution. In other words, the matrix D_T provides us with the right scaling to get asymptotic normality of $\partial g(\theta^0)/\partial\theta'(\tilde{\theta}_T - \theta^0)$. However, we can prove that the standard practice of Wald testing is valid even without Lee's assumption:

Theorem 4.1. *(Wald test)*
Under the assumptions 1 to 3 and if $g(.)$ is twice continuously differentiable, the Wald test statistic ζ_T^W for testing $H_0 : g(\theta) = 0$ is asymptotically $\chi^2(q)$ under the null.

While a detailed proof of theorem 4.1 is provided in the appendix, it is worth explaining why it works in spite of the aforementioned singularity problem. The key intuition is somewhat related to the well-known phenomenon that the finite sample performance of the Wald test depends on the way the null hypothesis is formulated[18].

Let us first imagine a fictitious situation where the range of $[\partial\rho_1'(\theta^0)/\partial\theta]$ is known. Then it is always possible to define a (q, q) nonsingular matrix H and a q dimensional function $h(\theta) = Hg(\theta)$ to ensure a *genuine disentangling* of the strongly identified and nearly-weakly identified directions to be tested. By genuine disentangling, we mean that for some q_1 such that $1 \leq q_1 \leq q$:

- for $j = 1, \cdots, q_1$: $[\partial h_j(\theta^0)/\partial\theta]$ belongs to $Im\,[\partial\rho_1'(\theta^0)/\partial\theta]$
- for $j = q_1 + 1, \cdots, q$: $[\partial h_j(\theta^0)/\partial\theta]$ does not belong to $Im\,[\partial\rho_1'(\theta^0)/\partial\theta]$ and no linear combinations of them do.

Then the perverse asymptotic singularity of example 4.1 is clearly avoided. Of course, at a deeper level, the new restrictions $h(\theta) = 0$ to be tested should be interpreted as a nonlinear transformation of the initial ones $g(\theta) = 0$ (since the matrix H depends on θ). It turns out that, for all practical purposes, by fictitiously seeing H as known, the Wald test statistics written with $h(.)$ or $g(.)$ are numerically equal. The proof of theorem 4.1 shows that this is the key reason why standard Wald test always works (despite appearing invalid at first sight).

As far as the size of the test is concerned, the existence of the two rates of convergence does not modify the standard Wald result. Of course, the power of the test heavily depends on the

[18] In some respect, our approach of nearly-weak identification complements the higher order expansions of Phillips and Park (1988).

strength of identification of the various constraints to test. More precisely, if, for the sake of notational simplicity, we consider only $q = 1$ restriction to test, we get:

Theorem 4.2. *(Local alternatives)*
Under assumptions 1 to 3, the Wald test of $H_0 : g(\theta) = 0$ (with $g(.)$ one dimensional continuously differentiable) is consistent under the sequence of local alternatives $H_{1T} : g(\theta) = 1/\delta_T$ if and only if either

$$\frac{\partial g(\theta^0)}{\partial \theta} \in Im\left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta}\right] \quad and \quad \delta_T = o(\sqrt{T})$$

or

$$\frac{\partial g(\theta^0)}{\partial \theta} \notin Im\left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta}\right] \quad and \quad \delta_T = o(\lambda_T)$$

The proof of theorem 4.2 is rather straightforward. In the line of the comments following theorem 3.1, a nonlinear function $g(.)$ of θ, interpreted as $\left[g(\theta^0) + \frac{\partial g(\theta^0)}{\partial \theta'}(\theta - \theta^0)\right]$, is identified at the standard rate $\sqrt{T}$ if and only if

$$\frac{\partial g(\theta^0)}{\partial \theta'} \in Im\left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta}\right]$$

5 Examples

We now work out several examples to illustrate the general theory of the previous sections as well as to shed some light on the link between our approach and Stock and Wright (2000).

5.1 Single-equation linear IV model

As already mentioned, the major difference between Stock and Wright's (2000) framework and ours lies in considering the subvector of strongly identified parameters as known a priori. The context of the linear IV regression model sheds some light on the relationships linking the two procedures. Consider the following single-equation linear IV model with two structural parameters, two orthogonal instruments and no exogenous variables for convenience:

$$\left\{\begin{array}{ccccccc} y & = & Y & \theta & + & u \\ (T,1) & & (T,2) & (2,1) & & (T,1) \\ Y & = & [X_1\ X_2] & C & + & [V_1\ V_2] \\ (T,2) & & (T,2) & (2,2) & & (T,2) \end{array}\right. \tag{5.1}$$

As commonly done in the literature the matrix of coefficients C is artificially linked to the sample size T in order to introduce some (nearly)-weak identification issues. However, to accommodate both interpretations of the identification issues, the matrices C_T are different. For our characterization (directly through the moment conditions) and for Stock and Wright's characterization (through the parameters) we have respectively:

$$C_T^{AR} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21}/T^{\lambda} & \pi_{22}/T^{\lambda} \end{bmatrix} \quad \text{and} \quad C_T^{SW} = \begin{bmatrix} \pi_{11} & \pi_{12}/T^{\lambda} \\ \pi_{21} & \pi_{22}/T^{\lambda} \end{bmatrix} \tag{5.2}$$

Choosing C_T^{AR} modifies the explanatory power of the second instrument X_2 only. As a result, one standard moment condition naturally emerges (associated with X_1) and one less informative (associated with X_2). Intuitively, the standard restriction should identify one standard direction in the parameter space, which is so far unknown. On the other hand, choosing C_T^{SW} is equivalent to modeling θ_2 as weakly identified. The price to pay for such an early knowledge is the alteration of the explanatory powers of both instruments. Typically the moment conditions,

$$E\left[(y_t - Y_t'\theta)X_t\right]$$

are respectively written as:

$$\left\{\begin{array}{ccccc} & E(X_{1t}^2)\pi_{11}(\theta_1^0 - \theta_1) & + & & E(X_{1t}^2)\pi_{12}(\theta_2^0 - \theta_2) \\ \frac{1}{T^{\lambda}} & E(X_{2t}^2)\pi_{21}(\theta_1^0 - \theta_1) & + & \frac{1}{T^{\lambda}} & E(X_{2t}^2)\pi_{22}(\theta_2^0 - \theta_2) \end{array}\right.$$

and

$$\left\{\begin{array}{ccccc} E(X_{1t}^2)\pi_{11}(\theta_1^0 - \theta_1) & + & \frac{1}{T^{\lambda}} & E(X_{1t}^2)\pi_{12}(\theta_2^0 - \theta_2) \\ E(X_{2t}^2)\pi_{21}(\theta_1^0 - \theta_1) & + & \frac{1}{T^{\lambda}} & E(X_{2t}^2)\pi_{22}(\theta_2^0 - \theta_2) \end{array}\right.$$

or, in a more compact way,

$$\left\{\begin{array}{l} \rho_1^{AR}(\theta_1, \theta_2) \\ \frac{1}{T^{\lambda}}\rho_2^{AR}(\theta_1, \theta_2) \end{array}\right. \quad \text{and} \quad \left\{\begin{array}{l} m_{1s}^{SW}(\theta_1) + \frac{1}{T^{\lambda}} m_{1w}^{SW}(\theta_2) \\ m_{2s}^{SW}(\theta_1) + \frac{1}{T^{\lambda}} m_{2w}^{SW}(\theta_2) \end{array}\right. \tag{5.3}$$

for some real functions $\rho_1^{AR}(.)$, $\rho_2^{AR}(.)$, $m_{1s}^{SW}(.)$, $m_{1w}^{SW}(.)$, $m_{2s}^{SW}(.)$ and $m_{2w}^{SW}(.)$.

We now introduce our reparametrization of section 2 to identify the standard direction in the parameter space. The derivative of the standard moment restriction is

$$J_1^0 = \frac{\partial \rho_1(\theta^0)}{\partial \theta'} = \left[-E(Y_{1t}X_{1t}) \vdots -E(Y_{2t}X_{1t}) \right] = \left[-E(X_{1t}^2)\pi_{11} \vdots -E(X_{1t}^2)\pi_{12} \right]$$

Hence the null space of J_1^0 is characterized by the following vector:

$$R = \begin{bmatrix} -\pi_{12} \\ \pi_{11} \end{bmatrix} \mu \qquad \text{where } \mu \in \mathbb{R}^*$$

It is then completed into a legitimate matrix of change of basis R^0 in the parameter space $\mathbb{R}^2$:

$$R^0 = \begin{bmatrix} a & -\pi_{12}\mu \\ b & \pi_{11}\mu \end{bmatrix} \qquad \text{with } (a,b) \in \mathbb{R}^2 \,/\, a\pi_{11} \neq -b\pi_{12}$$

The new parameter η can now be defined as follows: $\eta = [R^0]^{-1}\theta$ that is

$$\begin{cases} \eta_1 = \frac{1}{a\pi_{11}+b\pi_{12}} \left[\pi_{11}\theta_1 + \pi_{12}\theta_2 \right] \\ \eta_2 = -\frac{b}{\mu(a\pi_{11}+b\pi_{12})}\theta_1 + \frac{a}{\mu(a\pi_{11}+b\pi_{12})}\theta_2 \end{cases}$$

The standard direction is completely determined and not the weaker one. The main reason comes from the fact that everything that is not standard is weaker: in fact, the weaker directions contaminate the standard ones, when considering a linear combination of the two.

The above calculation shows that, strictly speaking, Stock and Wright (2000) and their linear reinterpretation of Staiger and Stock (1997) are not nested in our setting because each of their moment condition contains a strong part (that only depends on a subvector of parameter) and a weak part. Note that this setting (through the definition of the matrix C_T^{SW}) is conveniently built so as to know *a priori* which subset of the parameters is strongly identified. Now, if we pretend that we did not realize that the set of strongly identified parameter was known and we still perform the change of variables, we get:

$$J_1^0 \simeq [-\pi_{11}\ 0] \quad \text{hence} \quad R = [0\ \mu]' \quad \text{with} \quad \mu \in \mathbb{R}^*$$

and the change of basis is defined as:

$$R^0 = \begin{bmatrix} a & 0 \\ b & \mu \end{bmatrix} \quad \text{with } a \neq 0 \implies \eta = \begin{bmatrix} \mu & 0 \\ -b & a \end{bmatrix}$$

As expected, we identify the strongly identified direction as being parallel to θ_1. This is a nice internal consistency result of our procedure.

5.2 Non-linear IV model

As already mentioned in the linear case, our general setting does not strictly nest Stock and Wright (2000). However, we can show that the two procedures are relatively close to each other. Recall first the underlying assumptions on the moment restrictions:

$$Nearly-Weak \qquad E\left[\overline{\phi}_T(\theta)\right] = \frac{\Lambda_T}{\sqrt{T}}\rho(\theta)$$

$$Staiger-Stock \qquad E\left[\overline{\phi}_T(\theta)\right] = m_1(\theta_1) + \frac{1}{\sqrt{T}}m_2(\theta)$$

where θ_1 is the *a priori assumed* strongly identified parameter.

Let us now derive the first-order conditions associated with our minimization problem when the weighting matrix is chosen to be block diagonal such that $\Omega_D = diag[\Omega_{D1}\ \Omega_{D2}]$ with Ω_{Di} symmetric full rank (k_i, k_i)-matrix, i=1,2:

$$\min_{\theta}\left[\overline{\phi}_{1T}'(\theta)\Omega_{D1}\overline{\phi}_{1T}(\theta) + \overline{\phi}_{2T}'(\theta)\Omega_{D2}\overline{\phi}_{2T}(\theta)\right]$$

The associated first order conditions are

$$\frac{\partial\overline{\phi}_{1T}'(\hat{\theta}_T)}{\partial\theta}\Omega_{D1}\overline{\phi}_{1T}(\hat{\theta}_T) + \frac{\partial\overline{\phi}_{2T}'(\hat{\theta}_T)}{\partial\theta}\Omega_{D2}\overline{\phi}_{2T}(\hat{\theta}_T) = 0$$

The above first order condition can be seen as the selection of linear combinations of $\overline{\phi}_T$. If $\overline{\phi}_{1T}$ only depends on θ_1 then, after imposing $\lambda_T = 1$, our resulting linear combinations of the moment restrictions correspond to the ones of Stock and Wright (2000).

Note also that the null space used to reparametrize the problem can be defined directly from the above first order conditions after realizing that:

$$\left[\frac{\partial\overline{\phi}_{1T}'(\hat{\theta}_T)}{\partial\theta}\right]\Omega_{D1} \xrightarrow{P} \frac{\partial\rho_1'(\theta^0)}{\partial\theta}\Omega_{D1}$$

where $\Omega_{D1}\left[\partial\rho_1(\theta^0)/\partial\theta'\right]$ defines the same null space as $\left[\partial\rho_1(\theta^0)/\partial\theta'\right]$ since Ω_{D1} is a full rank squared matrix.

5.3 Estimation of the rates of convergence

In some special convenient cases (as a Monte Carlo study), it is possible to estimate the rate of convergence of our estimators via a linear regression. The idea is to simulate the model

for several sample sizes: for each sample size, the simulation is replicated M times to get M draws of the estimator. The Monte-Carlo distribution of the estimate can then be deduced and its variance calculated. Finally, the regression of logarithm of the variance on the constant regressor and the logarithm of the sample size is performed:

$$\log(Var(\hat{\theta}_T)) = \alpha + \beta \log T + u_T \tag{5.4}$$

where u_T is some error term. β can be estimated by OLS and it gives an estimate of the square of the convergence rate.

Section 6 below provides some illustrations of the estimation of the rates of convergence.

6 Monte-Carlo Study

6.1 Single-Equation linear IV model

In our first Monte-Carlo study, our goal is to verify the finite sample relevance of our asymptotic theory. In particular, we use the linear regression technique developed in section 5.3 to estimate the rates of convergence of the transformed parameters as well as the ones of the original parameters. Recall first the linear model of example 1 in section 3.2:

$$\begin{cases} \underset{(T,1)}{y} = \underset{(T,2)}{Y} \underset{(2,1)}{\theta} + \underset{(T,1)}{u} \\ \underset{(T,2)}{Y} = \underset{(T,2)}{[X_1\ X_2]} \underset{(2,2)}{C_T} + \underset{(T,2)}{[V_1\ V_2]} \end{cases} \tag{6.1}$$

$$\text{with} \quad C_T = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21}/T^{\mu} & \pi_{22}/T^{\mu} \end{bmatrix} \quad \text{and} \quad 0 < \mu < 1/2$$

The above model is estimated for several sample sizes as well as several degrees of weakness. We provide the results for $\mu = 1/5$: it corresponds to a *slow* convergence rate equal to $\lambda_T = T^{0.3}$, as introduced in section 2. This is a *strong nearly-weak* identification case and λ_T satisfies assumption 3.

Generally speaking the results are pretty good and conform to the theory. The main findings are listed here: i) The variance decreases with the sample size for the four estimators $\hat{\theta}_{1T}$,

$\hat{\theta}_{2T}$, $\hat{\eta}_{1T}$ and $\hat{\eta}_{2T}$. Moreover figure I.1 plots the log-variance as a function of the log-sample size: for the above estimators, it is a fairly straight decreasing line. This gives support to the fact that the variance is proportional to the sample size raised at some power;

ii) We now compare the rates of convergence among the sets of parameters by studying ratios of parameters, or specifically $\hat{\eta}_{2T}/\hat{\eta}_{1T}$ and $\hat{\theta}_{1T}/\hat{\theta}_{2T}$. From figure I.2, the ratio of the new set of parameters increase with the sample size whereas the ratio of the original set of parameters is fairly flat. This supports the fact that $\hat{\eta}_{1T}$ converges faster than $\hat{\eta}_{2T}$, whereas $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ converge at a similar rate;

iii) Finally, we present the results of the estimation of the rates of convergence with the linear regression technique described in section 5.3. See tables I.2 and I.3. According to our asymptotic theory, we expect to find a standard rate of $T^{1/2}$ for $\hat{\eta}_{1T}$ and a slow rate equal to $T^{0.3}$ for the three remaining parameters. Over the entire sample, the standard rate is relatively well estimated. On the other hand, the slow rate is less precise and we cannot conclude to the equality of the rates of convergence for $\hat{\eta}_{2T}$, $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$. However, when we consider only larger sample sizes (>5000), we get closer to the expected result. Since the convergence is slower, more data are expected to be needed to conclude.

6.2 CCAPM

In this section, we report some Monte-Carlo evidence about the intertemporally separable consumption capital asset pricing model (CCAPM) with constant relative risk-aversion (CRRA) preferences. The artificial data are generated in order to mimic the dynamic properties of the historical data. Hence, we can assess the empirical relevance of our general setting in such a context.

Moment conditions

The Euler equations lead to the following moment conditions:

$$E\left[h_{t+1}(\theta)|\mathcal{I}_t\right] = 0 \quad \text{with} \quad h_t(\theta) = \delta r_t c_t^{-\gamma} - 1$$

Our parameter of interest is then $\theta = [\delta \ \gamma]'$, with δ the discount factor and γ the preference parameter; (r_t, c_t) denote respectively a vector of asset returns and the consumption growth at time t. To be able to estimate this model, our K instruments $Z_t \in \mathcal{I}_t$ include the constant

as well as some lagged variables. We then rewrite the above moment conditions as

$$E_0\left[\phi_{t,T}(\theta)\right] = E_0\left[h_{t+1}(\theta) \otimes Z_{t,T}\right]$$

Note that to stress the potential weakness of the instruments (and as a result of the moment function, see section 2.1), we add the subscript T.

Data Generation:
Our Monte-Carlo design follows Tauchen (1986), Kocherlakota (1990), Hansen, Heaton and Yaron (1996) and more recently Stock and Wright (2000). More precisely, the artificial data are generated by the method discussed in Tauchen and Hussey (1991). This method fits a 16 state Markov chain to the law of motion of the consumption and the dividend growths, so as to approximate a beforehand calibrated gaussian VAR(1) model (see Kocherlakota (1990)). The CCAPM-CRRA model is then used to price the stocks and the riskfree bond in each time period, yielding a time series of asset returns.

It is important to stress that since the data are generated from a general equilibrium model, even the econometrician does not know whether (δ, γ) are (nearly)-weakly identified or not. In a similar study, Stock and Wright (2000) impose a different treatment for the parameters δ and γ: typically, δ is taken as strongly identified whereas γ is not. We do not make such an assumption. We are actually able, through a convenient reparametrization, to identify some directions of the parameter space that are strongly identified and some others that are not.

Strong and weak moment conditions:
We consider here three instruments: the constant, the centered lagged asset return and the centered lagged consumption growth. To be able to apply our nearly-weak GMM estimation, we need to separate the instruments (and the associated moment conditions) according to their *strength*. Typically, a moment restriction $E[\phi_t(\theta)]$ is (nearly)-weak when $E[\phi_t(\theta)]$ is *close* to zero for all θ. This means that the restriction does not permit to (partially) identify θ. Hence, we decide to evaluate each moment restriction for a grid of parameter values. If the moment is uniformly *close* to 0 then we conclude to its weakness. Note that this study can always be performed and is not specifically related to the Monte-Carlo setting; the Monte-Carlo setting is simply convenient to get rid of the simulation noise by averaging over the many simulated samples.

Figure I.3 has been built with a sample size of 100 and 2500 Monte-Carlo replications. Note that the conclusions are not affected when larger sample sizes are considered.

The above study clearly reveals two groups of moment restrictions: i) with the constant instrument, the associated restriction varies quite substantially with the parameter θ; ii) with the lagged instruments, both associated restrictions remain fairly small when θ vary over the grid. The set of instruments, and accordingly of moment conditions, is then separated as follows:

$$\phi_{tT}(\theta) = \begin{pmatrix} (\delta r_t c_t^{-\gamma} - 1) \\ (\delta r_t c_t^{-\gamma} - 1) \otimes \begin{bmatrix} r_{t-1} - \bar{r} \\ c_{t-1} - \bar{c} \end{bmatrix} \end{pmatrix}$$

Accordingly,

$$\bar{\phi}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \phi_{tT}(\theta) \quad \text{with} \quad \sqrt{T}E[\bar{\phi}_T(\theta)] = \begin{pmatrix} \lambda_{1T} & 0_{1,2} \\ 0_{2,1} & \lambda_{2T} Id_2 \end{pmatrix} \begin{pmatrix} \rho_1(\theta) \\ \rho_2(\theta) \end{pmatrix}$$

As emphasized earlier, our Monte-Carlo study simulates a general equilibrium model. So, even the econometrician does not know in advance which moment conditions are weak and the level of this weakness. Hence, λ_{1T} and λ_{2T} must be chosen so as to fulfill the following conditions, $\lambda_{1T} = o(\sqrt{T})$, $\lambda_{2T} = o(\lambda_{1T})$ and $\lambda_{1T} = o(\lambda_{2T}^2)$.

In their theoretical considerations (section 4.1), Stock and Wright (2000) also treat differently the covariances of the moment conditions. The strength of the constant instrument is actually used to provide some intuition on their identification assumptions (δ strongly identified and γ weakly identified). However, we maintain that if γ is weakly identified, then it affects the covariance between r_t and $c_t^{-\gamma}$, and hence the identification of δ is altered too. This actually matches some asymptotic results of Stock and Wright (2000) where the weak parameter affects the strong one, by preventing it to converge to a standard gaussian random variable.

We now identify the strong directions in the parameter space via the reparametrization introduced in section 3.

Reparametrization:

First, we define the matrix of the change of basis (or reparametrization), that enables us to identify the standard directions in the parameter space. Recall that it is defined through the null space of the following matrix,

$$J_1^0 = \frac{\partial \rho_1(\theta^0)}{\partial \theta'}$$

Straightforward calculations lead to:

$$\left[\frac{\partial \varphi_{1,t}(\theta)}{\partial \theta'}\right] = \left[\frac{\partial \varphi_{1,t}(\theta)}{\partial \delta} \quad \frac{\partial \varphi_{1,t}(\theta)}{\partial \gamma}\right] = \left[r_t c_t^{-\gamma} \;\vdots\; -\gamma\delta r_t c_t^{-\gamma-1}\right]$$

J_1^0 is then approximated as follows:

$$\bar{J} = \frac{\partial \bar{\rho}_1(\theta^0)}{\partial \theta'} = \left[\frac{1}{T}\sum_{t=1}^{T} r_t c_t^{-\gamma^0} \;\vdots\; -\frac{\gamma^0\delta^0}{T}\sum_{t=1}^{T} r_t c_t^{-\gamma^0-1}\right]$$

The null space of J_1^0 is defined via the (2,1)-matrix R_2 such that,

$$J_1^0 R_2 = 0 \Leftrightarrow R_2 = \nu \begin{bmatrix} -J_{12} \\ J_{11} \end{bmatrix} \quad \text{for any nonzero real number } \nu$$

R_2 is then completed with R_1 into the matrix R^0 so as to define a legitimate reparametrization. In other words, R^0 is of full rank. So practically the only constraint is the following,

$$R_1 = \begin{bmatrix} a \\ b \end{bmatrix} \quad \text{with} \quad \frac{a}{b} \neq -\frac{J_{12}}{J_{11}}$$

We then get,

$$R^0 = \begin{bmatrix} a & -\nu J_2 \\ b & \nu J_1 \end{bmatrix} \quad \text{and} \quad [R^0]^{-1} = \frac{1}{\mu(aJ_{11}+bJ_{12})}\begin{bmatrix} \mu J_{11} & \mu J_{12} \\ -b & a \end{bmatrix}$$

And the new set of parameter is then obtained as,

$$\eta = [R^0]^{-1}\theta \Leftrightarrow \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{aJ_{11}+bJ_{12}}(J_{11}\delta + J_{12}\gamma) \\ \frac{1}{\mu(aJ_{11}+bJ_{12})}(-b\delta + a\gamma) \end{pmatrix}$$

We can see that the standard direction η_1 is completely determined: that is the relative weights on δ and γ are known. As a convention, we normalize all vectors to unity and we also ensure that the subspaces defined respectively by the columns of R_2 and of R_1 are orthogonal.

Asymptotic result:

Recall first the adapted asymptotic convergence result:

$$\begin{bmatrix} \lambda_{1T}(\hat{\eta}_{1T} - \eta_1^0) \\ \lambda_{2T}(\hat{\eta}_{2T} - \eta_2^0) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, (J^{0\prime} S(\theta^0)^{-1} J^0)^{-1})$$

We now provide some details on the calculation of the above asymptotic variance. J^0 is defined as:

$$J^0 = \begin{bmatrix} \frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_1 & 0 \\ 0 & \frac{\rho_2(\theta^0)}{\partial \theta'} R_2 \end{bmatrix}$$

The approximation of J^0 is easily deduced from what has been done above.

By assumption $S(\theta^0)$ is block diagonal and defined as,

$$S(\theta^0) = \begin{bmatrix} Var(\sqrt{T}\bar{\rho}_{1T}(\theta^0)) & 0 \\ 0 & Var(\sqrt{T}\bar{\rho}_{2T}(\theta^0)) \end{bmatrix}$$

and a usual sample estimator is used.

Results:

We now provide the results of our Monte-Carlo study. Again, we consider three instruments, the constant, the lagged asset return and the lagged consumption growth, and two sets of parameter: set 1 (or model M1a as in Stock and Wright (2000)) where $\theta^{0\prime} = [.97\ 1.3]$; set 2 (or model M1b) where $\theta^{0\prime} = [1.139\ 13.7]$. Model M1b has previously been found to produce non-normal estimator distributions.

First, as we have seen in the previous section, the matrix of reparametrization is not known (even in our Monte-Carlo setting) and it is actually data dependent. We then investigate the variability of the true new parameter η^0. We found that even with small sample size ($T = 100$), the (estimated) true new parameter is really stable and does not depend much on the realization of the sample. For our two models, we find the following true new parameter:

Set 1: $\eta^0 = [-0.4015\ 1.5715]$; $[R^0]^{-1} = \begin{bmatrix} .6281 & -.7782 \\ .7782 & .6281 \end{bmatrix}$; $J_1^0 = [1.0321\ \ -1.2788]$

Set 2: $\eta^0 = [-13.5984\ 2.0176]$; $[R^0]^{-1} = \begin{bmatrix} .0642 & -.9979 \\ .9979 & .0642 \end{bmatrix}$; $J_1^0 = [.8986\ \ -13.9780]$

To estimate our models, we use the 2-step nearly-weak GMM and we produce the estimation results also for the intermediate 1-step estimator.

Note also that the optimization resolution is not affected by the rates of convergence (λ_{1T} and λ_{2T}).

Our findings are: i) All the estimators are consistent; ii) The variances of the estimators (for both $\hat{\eta}_T$ and $\hat{\theta}_T$) decrease to 0 with the sample size. The direct comparison between the variances of the parameter is not much of interest, this is rather the ratio that carries some information; iii) According to our asymptotic results, in case of nearly-weak identification, the asymptotic variance of the new parameter $\hat{\eta}_{1T}$ should decrease (a lot) faster with the sample size than the one of $\hat{\eta}_{2T}$. Figure I.4 investigates this feature by plotting the evolution of the ratio of the Monte-Carlo variance of $\hat{\eta}_{2T}$ and the Monte-Carlo variance of $\hat{\eta}_{1T}$ with the sample size.

For set 1, the ratio of variances is fairly constant: this suggests that the variances of both parameter $\hat{\eta}_{1T}$ and $\hat{\eta}_{2T}$ decrease at the same speed towards 0. This actually supports previous findings that this model presents less severe case of nonstandard behaviors. However, this does not support our study of the strength of the moments (see figure I.3) and the presence of plateaus for two of them. For set 2, the ratio of η slightly decreases with the sample size, while nothing like this can be observed for initial parameters. This provides some support to our asymptotic approach even though the difference between the identification issues in the two sets 1 and 2 is not very compelling from figure I.4 or from the estimation of the rates of convergence (tables I.4 and I.5). When studying the ratios, the slope is not significantly different from 0 for the new parameters and slightly positive for the original parameters. Similarly, for set 2, all rates are also close to each other (0.48), slightly slower than for set 1 and significantly different from the usual rate $T^{1/2}$. The slope of the ratio of new parameters is significantly positive whereas this is not the case for the original parameters.

7 Conclusion

In a GMM context, this essay proposes a general framework to account for potentially weak instruments. In contrast with existing literature, the weakness is directly related to the moment conditions (through the instruments) and not to the parameters. More precisely, we consider two groups of moment conditions: the standard one associated with the standard rate of convergence $\sqrt{T}$ and the nearly-weak one associated with the slower rate λ_T. This framework ensures that GMM estimators of all parameters are consistent, but at rates possibly slower than usual. We also characterize the validity of the standard testing approaches like Wald and GMM-LM tests. Moreover, we identify and estimate some $\sqrt{T}$-consistent di-

rections in the parameter space. Such results are practically relevant since the knowledge of the slower rate of convergence is not required.

For notational and expositional simplicity, we have chosen here to focus on two groups of moment conditions only. The extension to considering several degrees of weakness (think of a practitioner using several instruments of different informational qualities) is quite natural. Antoine and Renault (2007) specifically consider multiple groups of moment conditions associated with specific rates of convergence (which may actually be faster and/or slower than the standard rate $\sqrt{T}$). Note however that they do not explicitly consider any applications to identification issues, but rather applications in kernel, unit-root, extreme values or continuous time environments.

Appendix

Proofs of the main results

Proof of Theorem 2.1 *(Consistency)*:
The consistency of the minimum distance estimator $\hat{\theta}_T$ is a direct implication of the identification assumption 1 jointly with the following lemma:

Lemma A.

$$\|\rho(\hat{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

Proof of lemma A: From (2.2), the objective function is written as follows

$$Q_T(\theta) = \left[\frac{\Psi_T(\theta)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\theta)\right]' \Omega_T \left[\frac{\Psi_T(\theta)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\theta)\right] \quad \text{where } \Lambda_T = \begin{bmatrix} Id_{k_1} & 0 \\ 0 & \frac{\lambda_T}{\sqrt{T}} Id_{k_2} \end{bmatrix}$$

Since $\hat{\theta}_T$ is the minimizer of $Q(.)$ we have in particular:

$$\begin{aligned} & Q_T(\hat{\theta}_T) \leq Q(\theta^0) \\ \Longrightarrow \quad & \left[\frac{\Psi_T(\hat{\theta}_T)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\hat{\theta}_T)\right]' \Omega_T \left[\frac{\Psi_T(\hat{\theta}_T)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho(\hat{\theta}_T)\right] \leq \frac{\Psi_T'(\theta^0)}{\sqrt{T}} \Omega_T \frac{\Psi_T(\theta^0)}{\sqrt{T}} \end{aligned}$$

Denoting $d_T = \Psi_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) - \Psi_T'(\theta^0)\Omega_T\Psi_T(\theta^0)$, we get:

$$\left[\Lambda_T\rho(\hat{\theta}_T)\right]' \Omega_T \left[\Lambda_T\rho(\hat{\theta}_T)\right] + 2\left[\Lambda_T\rho(\hat{\theta}_T)\right]' \Omega_T\Psi_T(\hat{\theta}_T) + d_T \leq 0$$

Let μ_T be the smallest eigenvalue of Ω_T. The former inequality implies:

$$\mu_T\|\Lambda_T\rho(\hat{\theta}_T)\|^2 - 2\|\Lambda_T\rho(\hat{\theta}_T)\| \times \|\Omega_T\Psi_T(\hat{\theta}_T)\| + d_T \leq 0$$

In other words, $x_T = \|\Lambda_T\rho(\hat{\theta}_T)\|$ solves the inequality:

$$x_T^2 - \frac{2\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T} x_T + \frac{d_T}{\mu_T} \leq 0$$

and thus with

$$\Delta_T = \frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|^2}{\mu_T^2} - \frac{d_T}{\mu_T}$$

we have:

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} - \sqrt{\Delta_T} \leq x_T \leq \frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} + \sqrt{\Delta_T}$$

Since $x_T \geq \lambda_T \|\rho_T(\hat{\theta}_T)\|$ we want to show that $x_T = \mathcal{O}_P(1)$ that is

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \text{ and } \Delta_T = \mathcal{O}_P(1)$$

which amounts to show that:

$$\frac{\|\Omega_T \Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \text{ and } \frac{d_T}{\mu_T} = \mathcal{O}_P(1)$$

Note that since $det(\Omega_T) \xrightarrow{P} det(\Omega) > 0$ no subsequence of Ψ_T can converge in probability towards zero and thus we can assume (for T sufficiently large) that μ_T remains lower bounded away from zero with asymptotic probability one. Therefore, we just have to show that:

$$\|\Omega_T \Psi_T(\hat{\theta}_T)\| = \mathcal{O}_P(1) \text{ and } d_T = \mathcal{O}_P(1)$$

Since $trace(\Omega_T) \xrightarrow{P} trace(\Omega)$ and the sequence $trace(\Omega_T)$ is upper bounded in probability, so are all the eigenvalues of Ω_T. Therefore, the required boundedness in probability just results from our assumption 1(ii) ensuring that:

$$\sup_{\theta \in \Theta} \|\Psi_T(\theta)\| = \mathcal{O}_P(1)$$

The proof of lemma A is completed. Let us then deduce the weak consistency of $\hat{\theta}_T$ by a contradiction argument. If $\hat{\theta}_T$ is not consistent, there exists some positive ϵ such that:

$$P\left[\|\hat{\theta}_T - \theta^0\| > \epsilon\right]$$

does not converge to zero. Then we can define a subsequence $(\hat{\theta}_{T_n})_{n \in \mathbb{N}}$ such that, for some positive η:

$$P\left[\|\hat{\theta}_{T_n} - \theta^0\| > \epsilon\right] \geq \eta \quad \text{for } n \in \mathbb{N}$$

Let us denote

$$\alpha = \inf_{\|\theta - \theta^0\| > \epsilon} \|\rho(\theta)\| > 0 \quad \text{by assumption 1(i)}$$

Then for all $n \in \mathbb{N}$:

$$P\left[\|\rho(\hat{\theta}_{T_n})\| \geq \alpha\right] > 0$$

When considering the identification assumption 1(iii), this last inequality contradicts lemma A. This completes the proof of consistency. ■

Proof of Theorem 2.2 *(Score test)*:
The entire proof is written under the maintained null hypothesis that $\theta_0 = \theta^0$. The score statistic can be written as follows:

$$\begin{aligned} LM_T(\theta_0) &= T\bar{\phi}_T'(\theta_0)S_T^{-1}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta}S_T^{-1}\bar{\phi}_T(\theta_0) \\ &= \left(S_T^{-1/2}\Psi_T(\theta_0)\right)'S_T^{-1/2}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1} \\ &\quad\times\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta'}S_T'^{-1/2}\left(S_T^{-1/2}\Psi_T(\theta_0)\right) \end{aligned}$$

From assumption 1(ii) $S_T^{-1/2}\Psi_T(\theta_0)$ is asymptotically distributed as a gaussian process with mean 0 and identity covariance matrix. To be able to conclude, we only need to find an invertible matrix D_T and a full column rank matrix B such that

$$\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}D_T \xrightarrow{P} B$$

This would ensure that

$$S_T^{-1/2}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta}S_T^{-1}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta'}S_T'^{-1/2}$$

is a full rank p idempotent matrix and this leads to the desired result. Using assumption 2(iii), we call s_1 the rank of $[\partial\rho_1(\theta_0)/\partial\theta']$ and $(p-s_1)$ the one of $[\partial\rho_2(\theta_0)/\partial\theta']$. Define

$$D_T = \left[D_1 \;\; \frac{\sqrt{T}}{\lambda_T}D_2\right]$$

where D_1 and D_2 are respectively (p, s_1) and $(p, p-s_1)$ full column rank matrices such that $D_2'D_1 = 0$ and the range of D_1 is the range of $[\partial\rho_1(\theta_0)/\partial\theta']$. This ensures that D_T is invertible for every fixed sample size T. We now have:

$$\begin{aligned} \frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}D_T &= \begin{bmatrix} \frac{\partial\bar{\phi}_{1T}(\theta_0)}{\partial\theta'}D_1 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\bar{\phi}_{1T}(\theta_0)}{\partial\theta'}D_2 \\ \frac{\partial\bar{\phi}_{2T}(\theta_0)}{\partial\theta'}D_1 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\bar{\phi}_{2T}(\theta_0)}{\partial\theta'}D_2 \end{bmatrix} \\ &\xrightarrow{P} \begin{bmatrix} \frac{\partial\rho_1(\theta_0)}{\partial\theta'}D_1 & 0 \\ 0 & \frac{\partial\rho_2(\theta_0)}{\partial\theta'}D_2 \end{bmatrix} \equiv B \quad \text{which is of full column rank } p \end{aligned}$$

where the zero south-west and north-east blocks of B are deduced respectively from assumptions 2(iv) and (v). ■

Proof of Theorem 2.3 *(Rate of convergence)*:
From lemma A $\|\rho(\tilde{\theta}_T)\| = \|\rho(\tilde{\theta}_T) - \rho(\theta^0)\| = \mathcal{O}_P(1/\lambda_T)$ and by application of the mean-value theorem, for some $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0 component by component, we get:

$$\left\| \frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} \left(\hat{\theta}_T - \theta^0 \right) \right\| = \mathcal{O}_P \left(\frac{1}{\lambda_T} \right)$$

Note that, by a common abuse of notation, we omit to stress that $\tilde{\theta}_T$ actually depends on the component of $\rho(.)$. The key point is that since $\rho(.)$ is continuously differentiable and $\tilde{\theta}_T$, as $\tilde{\theta}_T$, converges in probability towards θ^0, we have:

$$\frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} \xrightarrow{P} \frac{\partial \rho(\theta^0)}{\partial \theta'}$$

and thus:

$$\frac{\partial \rho(\theta^0)}{\partial \theta'} \times \left(\hat{\theta}_T - \theta^0 \right) = z_T$$

with $\|z_T\| = \mathcal{O}_P(1/\lambda_T)$. Since $\partial \rho(\theta^0)/\partial \theta'$ is full column rank, we deduce that:

$$\left(\hat{\theta}_T - \theta^0 \right) = \left[\frac{\partial \rho'(\theta^0)}{\partial \theta} \frac{\partial \rho(\theta^0)}{\partial \theta'} \right]^{-1} \frac{\partial \rho'(\theta^0)}{\partial \theta} z_T$$

also fulfills:

$$\left\| \hat{\theta}_T - \theta^0 \right\| = \mathcal{O}_P \left(\frac{1}{\lambda_T} \right)$$

■

Proof of Theorem 3.1 *(Asymptotic Normality)*:
First we need a preliminary result which naturally extend the convergence towards J^0 in (3.2) when the true value θ^0 is replaced by some preliminary consistent estimator θ_T^*.

Lemma B. *Under assumptions 1 to 3, if θ_T^* is such that $\|\theta_T^* - \theta^0\| = \mathcal{O}_P(1/\lambda_T)$, then*

$$\sqrt{T} \frac{\partial \bar{\phi}_T(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \xrightarrow{P} J^0 \quad \textit{when } T \to \infty$$

Proof of Lemma B:

First note that

$$\sqrt{T}\frac{\partial\overline{\phi}_T(\theta_T^*)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1} = \begin{bmatrix} \frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'}R_1^0 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'}R_2^0 \\ \frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'}R_1^0 & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'}R_2^0 \end{bmatrix}$$

To get the results, we have to show the following:

$$\begin{aligned} i) \quad & \frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'} \xrightarrow{P} \frac{\partial\rho_1(\theta^0)}{\partial\theta'} \\ ii) \quad & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'} \xrightarrow{P} \frac{\partial\rho_2(\theta^0)}{\partial\theta'} \\ iii) \quad & \frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'}R_1^0 \xrightarrow{P} 0 \\ iv) \quad & \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{1T}(\theta_T^*)}{\partial\theta'}R_2^0 \xrightarrow{P} 0 \end{aligned}$$

i) From assumption 2(iv), we have: $\frac{\partial\overline{\phi}_{1T}(\theta^0)}{\partial\theta'} - \frac{\partial\rho_1(\theta^0)}{\partial\theta'} = o_P(1)$. The mean-value theorem applies to the k^{th} component of $[\partial\overline{\phi}_{1T}/\partial\theta']$ for $1 \leq k \leq k_1$. For some $\tilde{\theta}$ between θ^0 and θ_T^*, we have:

$$\frac{\partial\overline{\phi}_{1T,k}(\theta_T^*)}{\partial\theta'} - \frac{\partial\overline{\phi}_{1T,k}(\theta^0)}{\partial\theta'} = (\theta_T^* - \theta^0)'\frac{\partial^2\overline{\phi}_{1T,k}(\tilde{\theta}_T)}{\partial\theta\partial\theta'} = o_P(1)$$

where the last equality follows from assumption 3(ii) and the assumption on θ_T^*.

ii) From assumption 2(iv), we have:

$$\sqrt{T}\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'} - \lambda_T\frac{\partial\rho_2(\theta_0)}{\partial\theta'} = \mathcal{O}_P(1) \quad \Rightarrow \quad \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta^0)}{\partial\theta'} - \frac{\partial\rho_2(\theta^0)}{\partial\theta'} = o_P(1)$$

because $\lambda_T \xrightarrow{T} \infty$. The mean-value theorem applies to the kth component of $\partial\overline{\phi}_{2T}/\partial\theta'$ for $1 \leq k \leq k_2$. For some $\tilde{\theta}_T$ between θ^0 and θ_T^*, we have:

$$\frac{\sqrt{T}}{\lambda_T}\left(\frac{\partial\overline{\phi}_{2T,k}(\theta_T^*)}{\partial\theta'} - \frac{\partial\overline{\phi}_{2T,k}(\theta^0)}{\partial\theta'}\right) = (\theta_T^* - \theta^0)\frac{\sqrt{T}}{\lambda_T}\frac{\partial^2\overline{\phi}_{2T,k}(\tilde{\theta}_T)}{\partial\theta\partial\theta'} = o_P(1)$$

where the last equality follows from assumption 3(ii) and the assumption on θ_T^*.

iii)

$$\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'} = \frac{\lambda_T}{\sqrt{T}} \times \frac{\sqrt{T}}{\lambda_T}\frac{\partial\overline{\phi}_{2T}(\theta_T^*)}{\partial\theta'} = o_P(1)$$

because of (ii) and $\lambda_T = o(\sqrt{T})$.

iv) Recall the mean-value theorem from i). For $1 \leq k \leq k_1$ and $\tilde{\theta}_T$ between θ^0 and θ_T^*, we have:

$$\frac{\sqrt{T}}{\lambda_T}\frac{\partial \overline{\phi}_{1T,k}(\theta_T^*)}{\partial \theta'} = \frac{\sqrt{T}}{\lambda_T}\frac{\partial \overline{\phi}_{1T,k}(\theta^0)}{\partial \theta'} + \lambda_T(\theta_T^* - \theta^0)'\frac{1}{\lambda_T}\frac{\sqrt{T}}{\lambda_T}\frac{\partial^2 \overline{\phi}_{1T,k}(\tilde{\theta}_T)}{\partial\theta\partial\theta'}$$

The second member of the RHS is $o_P(1)$ because of assumptions 1(iii), 3(ii) and 3(iii) and the assumption on θ_T^*. Now we just need to show that the first member of the RHS is $o_P(1)$. Recall from assumption 2(v) that

$$\sqrt{T}\left[\frac{\partial \overline{\phi}_{1T}(\theta^0)}{\partial \theta'} - \frac{\partial \rho_1(\theta^0)}{\partial \theta'}\right] = \mathcal{O}_P(1) \Rightarrow \frac{\sqrt{T}}{\lambda_T}\left[\frac{\partial \overline{\phi}_{1T}(\theta^0)}{\partial \theta'} - \frac{\partial \rho_1(\theta^0)}{\partial \theta'}\right] R_2^0 = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right)$$

By definition R_2^0 is such that $\frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_2^0 = 0$. Hence we get

$$\frac{\sqrt{T}}{\lambda_T}\frac{\partial \overline{\phi}_{1T,k}(\theta^0)}{\partial \theta'} R_2^0 = \mathcal{O}_P\left(\frac{1}{\lambda_T}\right) = o_P(1)$$

This concludes the proof of lemma B. We now return to the proof of theorem 3.1. From the optimization problem (2.2), the first order conditions for $\hat{\theta}_T$ are written as:

$$\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta}\Omega \overline{\phi}_T(\hat{\theta}_T) = 0$$

A mean-value expansion yields to:

$$\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta}\Omega \overline{\phi}_T(\theta^0) + \frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta}\Omega \frac{\partial \overline{\phi}_T(\tilde{\theta}_T)}{\partial \theta'} \times \left(\hat{\theta}_T - \theta^0\right) = 0$$

where $\tilde{\theta}_T$ is between $\hat{\theta}_T$ and θ^0. Premultiplying the above equation by the non-singular matrix $T\tilde{\Lambda}_T^{-1}R^{0\prime}$ yields to an equivalent set of equations:

$$\hat{J}_T'\Omega\left[\sqrt{T}\overline{\phi}_T(\theta^0)\right] + \hat{J}_T'\Omega\tilde{J}_T \times \tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) = 0$$

after defining:

$$\hat{J}_T = \sqrt{T}\frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'}R^0\tilde{\Lambda}_T^{-1} \quad \text{and} \quad \tilde{J}_T = \sqrt{T}\frac{\partial \overline{\phi}_T(\tilde{\theta}_T)}{\partial \theta}R^0\tilde{\Lambda}_T^{-1}$$

From theorem 2.3 and lemma B, we can deduce that:

$$Plim\tilde{J}_T = J^0 \quad \text{and} \quad Plim\hat{J}_T = J^0$$

Hence,

$$\hat{J}_T'\Omega\hat{J}_T \xrightarrow{P} J^{0\prime}\Omega J^0 \quad \text{nonsingular by assumption}$$

Recall now that by assumption 1ii), $\Psi_T(\theta^0) = \sqrt{T}\left[\bar{\phi}_T(\theta^0)\right]$ converges to a normal distribution with mean 0 and variance $S(\theta^0)$. We then get the announced result.■

Proof of Theorem 3.2 *(Overidentifying test)*:
A Taylor expansion of order 1 of the moment conditions gives:

$$\begin{aligned}\sqrt{T}\bar{\phi}_T(\hat{\theta}_T) &= \sqrt{T}\bar{\phi}_T(\theta^0) + \sqrt{T}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}(\hat{\theta}_T - \theta^0) + o_P(1)\\ &= \sqrt{T}\bar{\phi}_T(\theta^0) + \hat{J}_T\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0) + o_P(1)\end{aligned}$$

with $\hat{J}_T = \sqrt{T}\partial\bar{\phi}_T(\hat{\theta}_T)/\partial\theta' R^0\tilde{\Lambda}_T^{-1}$.

A Taylor expansion of the FOC gives:

$$\begin{aligned}\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0) &= -\left[\left(\sqrt{T}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)' S_T^{-1}\left(\sqrt{T}\frac{\partial\bar{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)\right]^{-1}\\ &\quad\times\left(\sqrt{T}\frac{\partial\bar{\phi}_T(\hat{\theta})}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)' S_T^{-1}\sqrt{T}\bar{\phi}_T(\theta^0) + o_P(1)\end{aligned}$$

with S_T a consistent estimator of the asymptotic covariance matrix of the process $\Psi(\theta)$.

Combining the 2 above results leads to:

$$\sqrt{T}\bar{\phi}_T(\hat{\theta}_T) = \sqrt{T}\bar{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\bar{\phi}_T(\theta^0) + o_P(1)$$

Use the previous result to rewrite the criterion function:

$$\begin{aligned}TQ_T(\hat{\theta}_T) &= \left[\sqrt{T}\bar{\phi}_T(\hat{\theta}_T)\right]' S_T^{-1}\sqrt{T}\bar{\phi}_T(\hat{\theta}_T)\\ &= \left[\sqrt{T}\bar{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\bar{\phi}_T(\theta^0)\right]' S_T^{-1}\\ &\quad\times\left[\sqrt{T}\bar{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\bar{\phi}_T(\theta^0)\right] + o_P(1)\\ &= \left[\sqrt{T}\bar{\phi}_T(\theta^0)\right]' S_T^{-1}\sqrt{T}\bar{\phi}_T(\theta^0)\\ &\quad -\sqrt{T}\bar{\phi}_T(\theta^0)S_T^{-1}\hat{J}_T\left[\hat{J}_T'S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T'S_T^{-1}\sqrt{T}\bar{\phi}_T(\theta^0) + o_P(1)\\ &= \sqrt{T}\bar{\phi}_T(\theta^0)'S_T^{-1/2}\left[I - M\right]^{-1}S_T^{1/2}\sqrt{T}\bar{\phi}_T(\theta^0) + o_P(1)\end{aligned}$$

where $S_T^{1/2}$ is such that $S_T = S_T'^{1/2} S_T^{1/2}$ and $M = S_T^{-1/2} \hat{J}_T \left[\hat{J}_T' S_T^{-1} \hat{J}_T \right]^{-1} \hat{J}_T' S_T'^{-1/2}$ which is a projection matrix, hence idempotent and of rank $(K-p)$. The expected result follows. ■

Proof of Theorem 3.3 *(Orthogonalization)*:
Recall the inverse formulae:

$$\begin{aligned} S^{-1} &= \begin{bmatrix} [S_1^0]^{-1}(I + S_{12}^0 P^{-1} S_{21}^0 [S_1^0]^{-1}) & -[S_1^0]^{-1} S_{12}^0 P^{-1} \\ -P^{-1} S_{21} [S_1^0]^{-1} & P^{-1} \end{bmatrix} \\ &= \begin{bmatrix} Q^{-1} & -Q^{-1} S_{12}^0 [S_2^0]^{-1} \\ -[S_2^0]^{-1} S_{21}^0 Q^{-1} & S_2^{-1}(I + S_{21}^0 Q^{-1} S_{12}^0 [S_2^0]^{-1}) \end{bmatrix} \end{aligned}$$

with $Q = S_1^0 - S_{12}^0 [S_2^0]^{-1} S_{21}^0$ and $P = S_2^0 - S_{21}^0 [S_1^0]^{-1} S_{12}^0$.

Recall the block-diagonality of the matrix J^0 (see page 15):

$$J^0 = \begin{bmatrix} \frac{\partial \rho_1(\theta^0)}{\partial \theta'} R_1^0 & 0 \\ 0 & \frac{\partial \rho_2(\theta^0)}{\partial \theta'} R_2^0 \end{bmatrix}$$

Recall

$$AVar(\hat{\eta}_T) = \left[J^{0\prime} S^{-1} J^0 \right]^{-1} \quad \text{and} \quad AVar(\tilde{\eta}_T) = \left[J^{0\prime} [\Sigma^0]^{-1} J^0 \right]^{-1}$$

We need to compare the north-west and the south-east blocks of the above matrices.

Straightforward calculations lead to:

$$\left[J^{0\prime} [\Sigma^0]^{-1} J^0 \right]^{-1} = \begin{bmatrix} \left[\tilde{R}_1' Q^{-1} \tilde{R}_1 \right]^{-1} & 0 \\ 0 & \left[\tilde{R}_2' S_2^{-1} \tilde{R}_2 \right]^{-1} \end{bmatrix}$$

with $\tilde{R}_i \equiv \frac{\partial \rho_i}{\partial \theta'} R_i$ for $i = 1, 2$ and $Q \equiv S_1 - S_{12} S_2^{-1} S_{21}$.

On the other end, we have:

$$J^{0\prime} S^{-1} J^0 = \begin{bmatrix} \tilde{R}_1' Q^{-1} \tilde{R}_1 & -\tilde{R}_1' Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \\ -\tilde{R}_2' S_2^{-1} S_{21} Q^{-1} \tilde{R}_1 & \tilde{R}_2' \left[S_2^{-1} + S_2^{-1} S_{21} Q^{-1} S_{12} S_2^{-1} \right] \tilde{R}_2 \end{bmatrix} \equiv \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with $B' = C$.

i) We have $AVar(\hat{\eta}_{1T}) = A^{-1} + A^{-1} B (D - CA^{-1}B)^{-1} CA^{-1}$ that needs to be compared to $AVar(\tilde{\eta}_{1T}) = \left(\tilde{R}_1' Q^{-1} \tilde{R}_1 \right)^{-1}$.

Note that $A^{-1} = \left(\tilde{R}_1' Q^{-1} \tilde{R}_1\right)^{-1}$. Hence it is enough to study $A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$. Recall that $AVar(\tilde{\eta}_{2T}) = (D - CA^{-1}B)^{-1}$, hence it is a positive definite matrix (see also ii)). Also we have $B = C'$ and A symmetric. Then, we can deduce that $A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$ is positive semi-definite.

Finally, we can conclude: $AVar(\tilde{\eta}_{1T}) \geq\geq AVar(\tilde{\eta}_{1T})$

ii) We have $AVar(\tilde{\eta}_{2T}) = (D - CA^{-1}B)^{-1}$ that needs to be compared to $AVar(\tilde{\eta}_{2T}) = \left(\tilde{R}_2' S_2^{-1} \tilde{R}_2\right)^{-1}$.

It is enough to compare $D - CA^{-1}B$ with $\tilde{R}_2' S_2^{-1} \tilde{R}_2$.

$$\begin{aligned} D - CA^{-1}B &= \tilde{R}_2' S_2^{-1} \tilde{R}_2 + \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \\ &\quad - \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \end{aligned}$$

The last 2 terms of the RHS can be rewritten as follows:

$$\begin{aligned} & \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 - \tilde{R}_2' S_2^{-1} S_{21} Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} S_{12} S_2^{-1} \tilde{R}_2 \\ = \; & \tilde{R}_2' S_2^{-1} S_{21} \left\{ Q^{-1} - Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} \right\} S_{12} S_2^{-1} \tilde{R}_2 \end{aligned}$$

It is enough to study the middle matrix that appears between the brackets:

$$\begin{aligned} & Q^{-1} - Q^{-1} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1} \\ = \; & Q^{-1/2\prime} \left\{ I - Q^{-1/2} \tilde{R}_1 \left[\tilde{R}_1' Q^{-1/2\prime} Q^{-1/2} \tilde{R}_1\right]^{-1} \tilde{R}_1' Q^{-1/2\prime} \right\} Q^{-1/2} \\ = \; & Q^{-1/2\prime} \left\{ I - X(X'X)^{-1} X' \right\} Q^{-1/2} \\ = \; & Q^{-1/2\prime} M_X Q^{-1/2} \end{aligned}$$

with $Q^{-1} \equiv Q^{-1/2\prime} Q^{-1/2}$, $X \equiv Q^{-1/2} \tilde{R}_1$ and $M_X \equiv I - X(X'X)^{-1}X'$.

Finally, we have:

$$\begin{aligned} D - CA^{-1}B &= \tilde{R}_2' S_2^{-1} \tilde{R}_2 + \left(Q^{-1/2} S_{12} S_2^{-1} \tilde{R}_2\right)' M_X \left(Q^{-1/2} S_{12} S_2^{-1} \tilde{R}_2\right) \\ &\geq\geq \tilde{R}_2' S_2^{-1} \tilde{R}_2 \end{aligned}$$

because by definition, M_X is a projection matrix. Hence it is positive semi-definite as well as $H' M_X H$ for any matrix H.

We can then conclude: $AVar(\tilde{\eta}_{2T}) \leq\leq AVar(\tilde{\eta}_{2T})$. ■

Proof of Proposition 3.4:

We consider the following set of moment conditions,

$$\begin{pmatrix} \bar{\phi}^H_{1T}(\theta^0) \\ \bar{\phi}^H_{2T}(\theta^0) \end{pmatrix} = \begin{pmatrix} \bar{\phi}_{1T}(\theta^0) + H\bar{\phi}_{2T}(\theta^0) \\ \bar{\phi}_{2T}(\theta^0) \end{pmatrix}$$

such that

$$\sqrt{T}\begin{bmatrix} \phi^H_{1T}(\theta^0) - \rho_1(\theta^0) \\ \phi^H_{2T}(\theta^0) - \frac{\lambda_T}{\sqrt{T}}\rho_2(\theta^0) \end{bmatrix} \Rightarrow \tilde{\Psi}(\theta^0)$$

where $\tilde{\Psi}(\theta)$ is a Gaussian random variable with mean zero and variance

$$S^H = \begin{bmatrix} S^0_1 + HS^0_2H' + S^0_{12}H' + HS^0_{21} & S^0_{12} + HS^0_2 \\ S^0_{21} + S^0_2H' & S^0_2 \end{bmatrix}$$

The above set is orthogonalized as follows:

$$\begin{pmatrix} \tilde{\phi}^H_{1T}(\theta^0) \\ \tilde{\phi}^H_{2T}(\theta^0) \end{pmatrix} = \begin{pmatrix} \bar{\phi}^H_{1T}(\theta^0) - Cov\left(\sqrt{T}\bar{\phi}^H_{1T}(\theta^0), \sqrt{T}\bar{\phi}^H_{2T}(\theta^0)\right)\left[Var\left(\sqrt{T}\bar{\phi}^H_T(\theta^0)\right)\right]^{-1}\bar{\phi}^H_{2T}(\theta^0) \\ \bar{\phi}^H_{2T}(\theta^0) \end{pmatrix}$$

with

$$\begin{aligned} \tilde{\phi}^H_{1T}(\theta^0) &= \bar{\phi}_{1T}(\theta^0) + H\bar{\phi}_{2T}(\theta^0) - (S_{12} + HS_2)S_2^{-1}\bar{\phi}_{2T}(\theta^0) \\ &= \bar{\phi}_{1T}(\theta^0) + H\bar{\phi}_{2T}(\theta^0) - S_{12}S_2^{-1}\bar{\phi}_{2T}(\theta^0) - H\bar{\phi}_{2T}(\theta^0) \\ &= \tilde{\phi}_{1T}(\theta^0) \end{aligned}$$

■

Proof of Corollary 3.5:

The proof directly follows from the results of Theorem 3.3 and Property 3.4. ■

Proof of Lemma 3.6

In theorem 3.1, we have established that the following vector is asymptotically normally distributed:

$$\begin{bmatrix} \sqrt{T}.A\left(\hat{\theta}_T - \theta^0\right) \\ \lambda_T B\left(\hat{\theta}_T - \theta^0\right) \end{bmatrix}$$

We now show that the above convergence result is not altered when matrices A and B are replaced by some λ_T-consistent estimators, respectively $\hat{A}$ and $\hat{B}$.

(i) Convergence of the nearly-weak directions:

$$\lambda_T \hat{B}\left(\hat{\theta}_T - \theta^0\right) = \underbrace{\lambda_T B\left(\hat{\theta}_T - \theta^0\right)}_{(1)} + \underbrace{\lambda_T\left(\hat{B} - B\right)\left(\hat{\theta}_T - \theta^0\right)}_{(2)}$$

$(1) = \mathcal{O}_P(1)$. $\hat{B}$ is a λ_T-consistent estimator of B, so clearly (1) dominates (2): this is denoted as $(2) \prec (1)$.

(ii) Convergence of the standard directions:

$$\sqrt{T}\hat{A}\left(\hat{\theta} - \theta^0\right) = \underbrace{\sqrt{T}A\left(\hat{\theta} - \theta^0\right)}_{(1)} + \underbrace{\frac{\sqrt{T}}{\lambda_T}\left(\hat{A} - A\right)\lambda_T\left(\hat{\theta} - \theta^0\right)}_{(2)}$$

We have $(1) = \mathcal{O}_P(1)$ and $\lambda_T(\hat{\theta} - \theta^0) = \mathcal{O}_P(1)$. Hence,

$$(2) \prec (1) \iff \frac{\sqrt{T}}{\lambda_T}(\hat{A} - A) = o_P(1) \iff \hat{A} - A = o_P\left(\frac{\lambda_T}{\sqrt{T}}\right)$$

By assumption $\left\|\hat{A} - A\right\| = \mathcal{O}_P(\frac{1}{\lambda_T})$, so we get:

$$(2) \prec (1) \iff \frac{1}{\lambda_T} = o\left(\frac{\lambda_T}{\sqrt{T}}\right) \iff \sqrt{T} = o\left(\lambda_T^2\right)$$

which corresponds to assumption 3(i). ■

Proof of Theorem 4.1 *(Wald test)*:
The proof is divided into two steps:
- step 1: we define an algebraically equivalent formulation of $H_0 : g(\theta) = 0$ as $H_0 : h(\theta) = 0$ such that its first components are strongly identified while the remaining ones are nearly-weakly identified without any linear combinations of the latter being strongly identified.
- step 2: we show that the Wald test statistic on $H_0 : h(\theta) = 0$ asymptotically converges to the proper $\chi^2(q)$ distribution and that it is numerically equal to the Wald test statistic on $H_0 : g(\theta) = 0$.

- Step 1: The space of strongly identified directions to be tested is:

$$I^0(g) = \left[Im\frac{\partial g'(\theta^0)}{\partial \theta}\right] \cap \left[Im\frac{\partial \rho_1'(\theta^0)}{\partial \theta}\right]$$

Denote $n^0(g)$ the dimension of $I^0(g)$. Then, among the q restrictions to be tested, $n^0(g)$ are strongly identified and the $(q - n^0(g))$ remaining ones are nearly-weakly identified.
Define q vectors of $\mathbb{R}^q$ denoted as ϵ_j $(j = 1, \cdots, q)$ such that $[\partial g'(\theta^0)/\partial\theta \times \epsilon_j]_{j=1}^{q_1}$ is a basis of $I^0(g)$ and $[\partial g'(\theta^0)/\partial\theta \times \epsilon_j]_{j=q_1+1}^{q}$ is a basis of

$$\left[I^0(g)\right]^{\perp} \cap \left[Im \frac{\partial g'(\theta^0)}{\partial\theta}\right]$$

We can then define a new formulation of the null hypothesis $H_0 : g(\theta) = 0$ as: $H_0 : h(\theta) = 0$ where $h(\theta) = Hg(\theta)$ with H invertible matrix such that $H' = [\epsilon_1 \cdots \epsilon_q]$. The two formulations are algebraically equivalent since $h(\theta) = 0 \iff g(\theta) = 0$. Moreover,

$$\underset{T\to\infty}{Plim} \left[D_T \frac{\partial h(\theta^0)}{\partial\theta'} R^0 \left[\tilde{\Lambda}_T\right]^{-1}\right] = B^0$$

with D_T a (q, q) invertible diagonal matrix with its first $n^0(g)$ coefficients equal to $\sqrt{T}$ and the $(p - n^0(g))$ remaining ones equal to λ_T and B^0 a (q, p) matrix with full column rank.

- Step 2: first we show that the 2 induced Wald test statistics are numerically equal.

$$\begin{aligned}
\zeta_T^W(g) &= Tg'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial\theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial\theta} \right\}^{-1} g(\hat{\theta}_T) \\
&= TH'g'(\hat{\theta}_T) \left\{ H \frac{\partial g(\hat{\theta}_T)}{\partial\theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial\theta} H' \right\}^{-1} Hg(\hat{\theta}_T) \\
&= \zeta_T^W(h)
\end{aligned}$$

Then we show $\zeta_T^W(h) \xrightarrow{d} \chi^2(q)$. First we need a preliminary result which naturally extends the above convergence towards B^0 when θ^0 is replaced by a λ_T-consistent estimator θ_T^*:

$$Plim_T \left[D_T \frac{\partial h(\theta_T^*)}{\partial\eta'} [\tilde{\Lambda}_T]^{-1}\right] = B^0$$

The proof is very similar to lemma B in the proof of theorem 3.1 and is not reproduced here. Note that the fact that $g(.)$ is twice continuously differentiable is needed for this proof.

The Wald test statistic on $h(.)$ can be written as follows:

$$\begin{aligned}\zeta_T^W(h) &= T\left[D_T h(\hat{\theta}_T)\right]' \left\{ D_T \frac{\partial h(\hat{\theta}_T)}{\partial \theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial h'(\hat{\theta}_T)}{\partial \theta} D_T \right\}^{-1} \left[D_T h(\hat{\theta}_T)\right] \\ &= \left[D_T h(\hat{\theta}_T)\right]' \left\{ D_T \frac{\partial h(\hat{\theta}_T)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \left[\hat{J}_T' S_T^{-1} \hat{J}_T\right]^{-1} \tilde{\Lambda}_T^{-1} R^{0\prime} \frac{\partial h'(\hat{\theta}_T)}{\partial \theta} D_T \right\}^{-1} \left[D_T h(\hat{\theta}_T)\right]\end{aligned}$$

where $\hat{J}_T \equiv \sqrt{T} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1}$ with $\hat{J}_T \xrightarrow{P} J^0$ and $\hat{J}_T' S_T^{-1} \hat{J}_T \xrightarrow{P} J^{0\prime} \left[S(\theta^0)\right]^{-1} J^0 \equiv \Sigma$.
Now from the mean-value theorem under H_0 we deduce:

$$D_T h(\hat{\theta}_T) = D_T \frac{\partial h(\theta_T^*)}{\partial \theta'} \left(\hat{\theta}_T - \theta^0\right) = \left[D_T \frac{\partial h(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1}\right] \tilde{\Lambda}_T \left[R^0\right]^{-1} \left(\hat{\theta}_T - \theta^0\right)$$

$$\text{with } \left[D_T \frac{\partial h(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1}\right] \xrightarrow{P} B^0 \text{ and } \tilde{\Lambda}_T \left[R^0\right]^{-1} \left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

Finally we get

$$\xi_T^W(h) = \left[\tilde{\Lambda}_T \left[R^0\right]^{-1} (\hat{\theta}_T - \theta^0)\right]' B_0' (B_0 \Sigma B_0')^{-1} B_0 \left[\tilde{\Lambda}_T \left[R^0\right]^{-1} (\hat{\theta}_T - \theta^0)\right] + o_P(1)$$

Following the proof of theorem 3.2 we get the expected result. ■

T	Var($\hat{\eta}_{1T}$)	Var($\hat{\eta}_{2T}$)	Var($\hat{\theta}_{1T}$)	Var($\hat{\theta}_{2T}$)
50	0.0534	0.0264	0.0233	0.0081
100	0.0236	0.01799	0.0154	0.0056
200	0.0133	0.0103	0.0082	0.0037
300	0.0093	0.0080	0.0059	0.0030
400	0.0073	0.0058	0.0041	0.0024
500	0.0060	0.0049	0.0031	0.0021
600	0.0051	0.0044	0.0029	0.0019
700	0.0042	0.0042	0.0027	0.0018
800	0.0035	0.0039	0.0026	0.0017
900	0.0031	0.0035	0.0023	0.0015
1000	0.0028	0.0033	0.0022	0.0014
1500	0.0019	0.0023	0.0015	0.0010
2000	0.0014	0.0020	0.0012	0.0009
3000	0.0009	0.0015	0.0009	0.0007
5000	0.0005	0.0011	0.0006	0.0005
6000	0.0005	0.0010	0.0006	0.0005
7000	0.0004	0.0009	0.0005	0.0004
8000	0.0003	0.0008	0.0005	0.0004
9000	0.0003	0.0008	0.0004	0.0004
10000	0.0003	0.0007	0.0004	0.0003
11000	0.0002	0.0007	0.0004	0.0003
12000	0.0002	0.0006	0.0004	0.0003
13000	0.0002	0.0006	0.0003	0.0003

Table I.1: Single-equation linear IV model: Estimation results for the variance of the Monte Carlo distributions of the new parameters $\hat{\eta}_T$ as well as the original one $\hat{\theta}_T$ for various sample sizes.

Entire specter of sample sizes				
	$\hat{\beta}$	95% Confidence Interval		Estimated rate
$\hat{\eta}_{1T}$	-0.9976	-1.0111	-0.9842	0.4988
$\hat{\eta}_{2T}$	-0.6806	-0.6950	-0.6663	0.3403
$\hat{\theta}_{1T}$	-0.7577	-0.7808	-0.7345	0.3788
$\hat{\theta}_{2T}$	-0.6130	-0.6221	-0.6039	0.3065
Large sample sizes (>5000)				
	$\hat{\beta}$	95% Confidence Interval		Estimated rate
$\hat{\eta}_{1T}$	-0.9903	-1.0042	-0.9764	0.4951
$\hat{\eta}_{2T}$	-0.6267	-0.6692	-0.5841	0.3133
$\hat{\theta}_{1T}$	-0.6667	-0.7088	-0.6246	0.3333
$\hat{\theta}_{2T}$	-0.6046	-0.6411	-0.5681	0.3023

Table I.2: Single-equation linear IV model: Estimation of the β coefficients in the linear regression (5.4) and the rates of convergence of the variance series.

Entire specter of sample sizes			
	Estimated slope	95% Confidence Interval	
$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	0.3170	0.2905	0.3435
$Var(\hat{\theta}_{2T})/Var(\hat{\theta}_{1T})$	0.1447	0.1209	0.1685
Large sizes (>5000)			
	Estimated slope	95% Confidence Interval	
$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	0.3636	0.3114	0.4158
$Var(\hat{\theta}_{2T})/Var(\hat{\theta}_{1T})$	0.0621	0.0478	0.0764

Table I.3: Single-equation linear IV model: Estimation of the β coefficients for the ratio series.

Large sample sizes (>10000)								
	$\hat{\beta}$	95% CI		Rate		Slope	95% CI	
$\hat{\eta}_{1T}$	-0.9862	-1.0069	-0.9654	0.4931	$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	-0.0011	-0.0068	0.0046
$\hat{\eta}_{2T}$	-0.9872	-1.0071	-0.9674	0.4936				
$\hat{\theta}_{1T}$	-0.9879	-1.0072	-0.9686	0.4940	$Var(\hat{\theta}_{1T})/Var(\hat{\theta}_{2T})$	0.0006	0.0002	0.0010
$\hat{\theta}_{2T}$	-0.9885	-1.0077	-0.9693	0.4942				

Table I.4: CCAPM for set 1: i) Estimation of the β coefficients in the linear regression (5.4) and the rates of convergence of the variance series; ii) Estimation of the β coefficient for the ratio series

Large sample sizes (>10000)								
	$\hat{\beta}$	95% CI		Rate		Slope	95% CI	
$\hat{\eta}_{1T}$	-0.9674	-0.9872	-0.9477	0.4837	$Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$	0.0018	0.0010	0.0027
$\hat{\eta}_{2T}$	-0.9656	-0.9854	-0.9458	0.4828				
$\hat{\theta}_{1T}$	-0.9633	-0.9831	-0.9436	0.4816	$Var(\hat{\theta}_{1T})/Var(\hat{\theta}_{2T})$	-0.0002	-0.0052	0.0048
$\hat{\theta}_{2T}$	-0.9631	-0.9828	-0.9435	0.4815				

Table I.5: CCAPM for set 2: i) Estimation of the β coefficient in the linear regression (5.4) and the rates of convergence of the variance series; ii) Estimation of the β coefficient for the ratio series

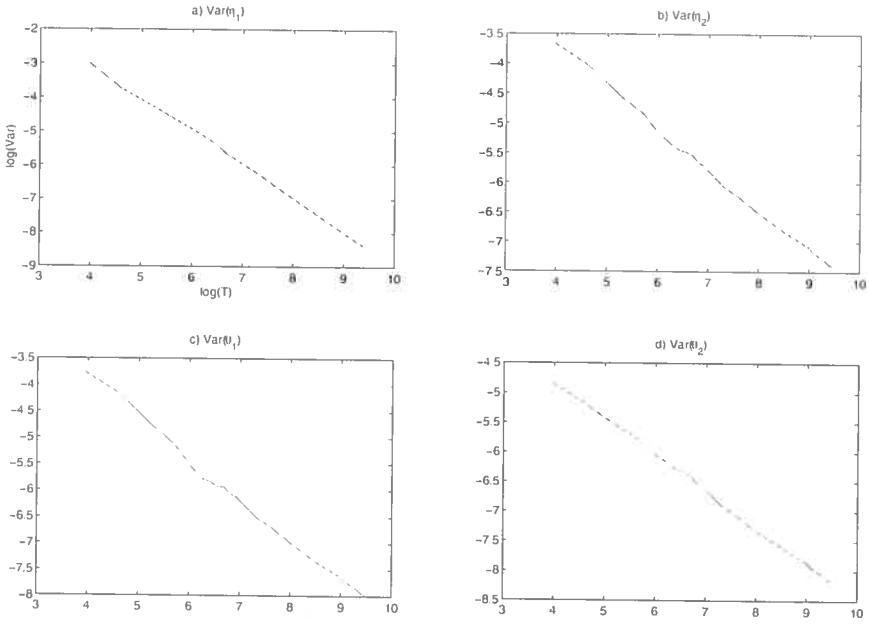

Figure I.1: Single-equation linear IV model: Logarithm of the variance as a function of the log-sample size. Top figures for the new parameters with a) $\hat{\eta}_{1T}$; b) $\hat{\eta}_{2T}$; Bottom figures for the original parameters with c) $\hat{\theta}_{1T}$; d) $\hat{\theta}_{2T}$.

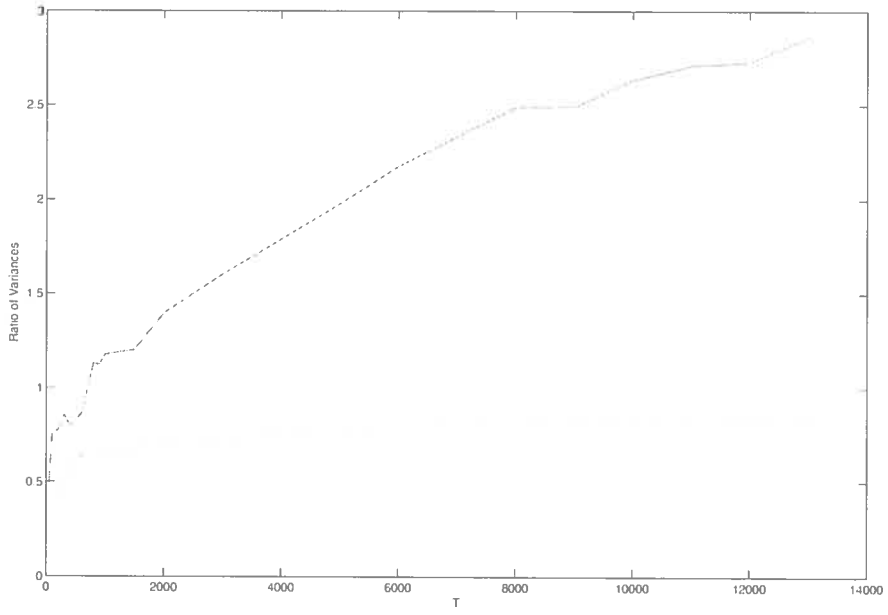

Figure I.2: Single-equation linear IV model: Ratio of the variance of the parameters as a function of the sample size. Solid line for $Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$: Dashed line for $Var(\hat{\theta}_{2T})/Var(\hat{\theta}_{1T})$.

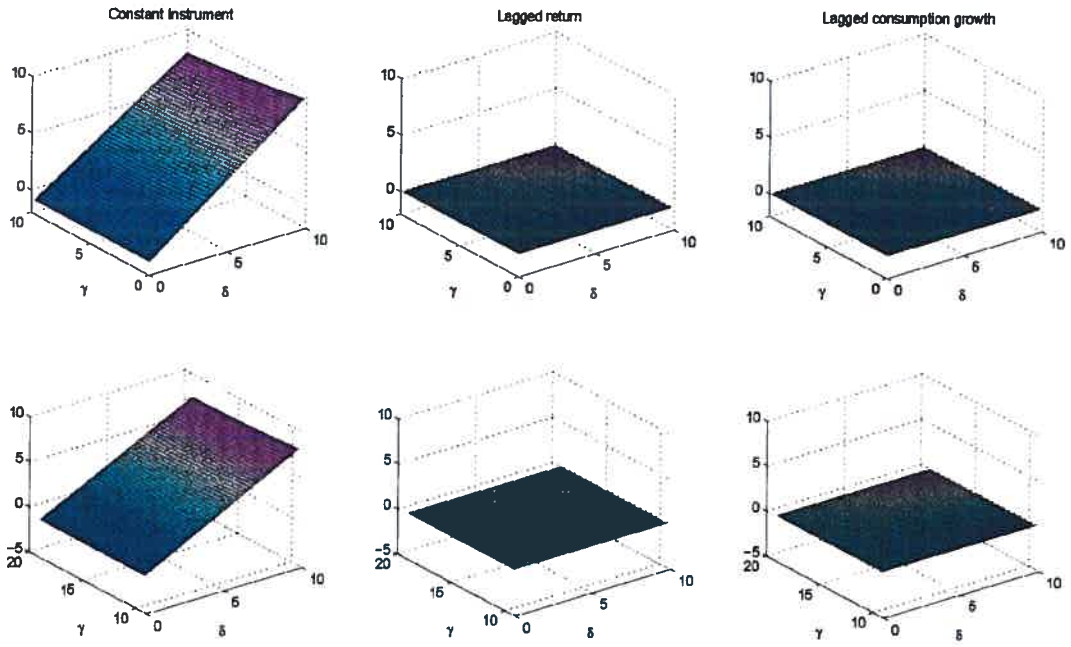

Figure I.3: CCAPM: Moment restrictions as a function of the parameter values θ. Top figures for set 1 with a) constant instrument; b) lagged asset return; c) lagged consumption rate. Bottom figures for set 2. T=100 and M=2500.

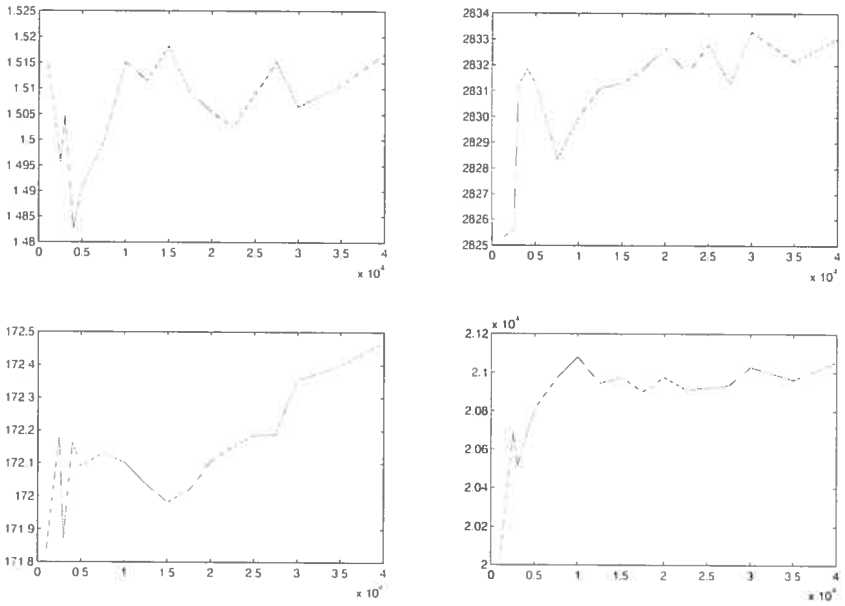

Figure I.4: CCAPM: Ratio of the variances as a function of the sample size. Top for set 1, left for $Var(\hat{\eta}_{2T})/Var(\hat{\eta}_{1T})$ and right for $Var(\hat{\theta}_{1T})/Var(\hat{\theta}_{2T})$. Bottom for set 2.

Chapitre II

Testing parameters in GMM without assuming that they are identified: a comment†

†This chapter is based on a paper co-authored with Eric Renault

1 Introduction

In a recent paper published in *Econometrica*, Kleibergen (2005) proposes a GMM-LM based statistic, the K statistic. It uses a modified estimator of the Jacobian, asymptotically uncorrelated with the empirical mean of the moments. This property permits to relax the full rank assumption on the Jacobian and even allows the application of the test in case of weak instruments. In this chapter, we shed some light on power calculations for the K and LM (or score) test statistics. These calculations are produced for several identification issues, from strong to weak, and for some mixture cases of the former.

Kleibergen (2005) starts with a joint central limit theorem on the moment conditions $\overline{\phi}_T(\theta^0)$ and their associated Jacobian $\left[\partial\overline{\phi}_T(\theta^0)/\partial\theta'\right]$:

Assumption 1. *(Joint CLT from Kleibergen (2005))*

$$\sqrt{T}\begin{bmatrix} \overline{\phi}_T(\theta^0) & - & E\left[\overline{\phi}_T(\theta^0)\right] \\ Vec\left[\frac{\partial\overline{\phi}_T(\theta^0)}{\partial\theta'}\right] & - & Vec\left[J(\theta^0)\right] \end{bmatrix} \quad \text{with } J(\theta^0) \equiv E\left[\frac{\partial\overline{\phi}_T(\theta^0)}{\partial\theta'}\right]$$

follows an asymptotic gaussian distribution with mean 0 and variance V.

The identification is strong when $J(\theta^0)$ is full-column rank and weak when there exists a deterministic matrix C such that $J(\theta^0) = \frac{C}{\sqrt{T}}$. To test $H_0 : \theta = \theta_0$, Kleibergen proposes the K-test, robust to any case where strong identification fails[1]. It is a modification of the standard GMM score test: instead of computing the LM test as a norm of $\left[\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\Omega^{-1}\overline{\phi}_T\right]$, $\left[\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]$ is replaced by the residual of its regression on the moment conditions. More formally, $\left[\frac{\partial\overline{\phi}_{iT}(\theta_0)}{\partial\theta_j}\right]$ is replaced by

$$\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} = \frac{\partial\overline{\phi}_{iT}(\theta_0)}{\partial\theta_j} - Cov\left[\sqrt{T}\frac{\partial\overline{\phi}_{iT}(\theta_0)}{\partial\theta_j}, \sqrt{T}\overline{\phi}(\theta_0)\right]\left[Var\left(\sqrt{T}\overline{\phi}(\theta_0)\right)\right]^{-1}\overline{\phi}(\theta_0)$$

It has been shown in the literature that this correction generally provides finite sample improvement, without modifying the standard first-order asymptotics: see e.g. Antoine, Bonnal and Renault (2007), Donald and Newey (2000) and Newey and Smith (2004).

[1] The precise identification pattern does not need to be known for the test to be valid and performed. Note also that we distinguish between the true value of the parameter (θ^0) and its value under the null hypothesis (θ_0).

We want to investigate the power of the K-test and compare it to the power of the standard score test. For this, we go one step further in the specification of the identification issues. We think that rank deficiencies of the Jacobian must be more tightly related to the moment conditions themselves. More precisely, we use the framework of chapter 1. Everything starts at the moment conditions level: they are partitioned according to the information they carry, say strong or nearly-weak. In this framework, the Jacobian naturally inherits a similar pattern, which may explain the asymptotic rank deficiencies. Since the knowledge or the estimation of the degree of weakness of each moment conditions is not required to perform inference, we find that this framework is not much more involved than Kleibergen (2005). Moreover, it helps clarifying power calculations as shown later.

The chapter is organized as follows. First, we quickly recall the framework of chapter 1. Then, we present the power calculations of the LM and K test statistics against a sequence of local alternatives. We also discuss testing subsets of parameters. Finally, we conclude.

All the proofs are gathered in the appendix.

2 Power against a sequence of local alternatives

2.1 Framework

In chapter 1, we proposed a framework where the moment conditions are partitioned in terms of the information they carry. Let us consider here similarly two groups of moment conditions and the associated central limit theorem assumption:

Assumption 2. *(CLT from Chapter 1)*

$$\sqrt{T}\begin{bmatrix} \overline{\phi}_{1T}(\theta^0) - \rho_1(\theta^0) \\ \overline{\phi}_{2T}(\theta^0) - \frac{\lambda_T}{\sqrt{T}}\rho_2(\theta^0) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, S(\theta^0)) \quad \textit{with} \quad 0 << \lambda_T << \sqrt{T}$$

The first group has k_1 standard moment conditions whereas the second one has k_2 weaker moment conditions. λ_T represents the degree of weakness of the second group of moment conditions, or the speed at which the associated information disappears. This is a convenient way to acknowledge that moment conditions may carry information of heterogeneous quality.

Weaker moment conditions contain fragile information that needs to be preserved because it is still relevant. We will see that, with heterogeneous quality of information, the transformation of Kleibergen may alter the asymptotic behavior of the test statistic. This is in contrast with standard GMM.

The Jacobian matrix naturally inherits the above special design:

Assumption 3. *(Assumption 2(iv) and 2(v) from Chapter 1)*

$$2(iv) \quad \left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta} \; \frac{\partial \rho_2'(\theta^0)}{\partial \theta}\right] = \underset{T\to\infty}{Plim} \left[\frac{\partial \overline{\phi}_{1T}'(\theta^0)}{\partial \theta} \; \frac{\sqrt{T}}{\lambda_T} \frac{\partial \overline{\phi}_{2T}'(\theta^0)}{\partial \theta}\right]$$

$$2(v) \quad \sqrt{T}\left[\frac{\partial \overline{\phi}_{1T}'(\theta^0)}{\partial \theta} - \frac{\partial \rho_1'(\theta^0)}{\partial \theta}\right] = \mathcal{O}_P(1)$$

2.2 Power of the K-test

We investigate the power of the LM and the K-test. Basically, if H_0, say $\theta = \theta_0$, is false, we would like to know the probability that it will be rejected. Since we work with asymptotic distributions, for any $\theta \neq \theta_0$, the answer is 1 with a consistent test: this does not help the comparison. Hence, instead of looking at an infinite sample, we want to find an approximation for the case of a finite (but reasonably large) sample. The classical solution is to assume that the data-generating process is subject to a Pitman drift. More precisely, the data in a sample of size T are generated by the model element $\theta^{(T)} = \theta_0 + \frac{\gamma}{\delta_T}$ with γ the direction and δ_T the rate of local departure. This device of using a sequence of local alternatives will be the basis of the following discussion of power properties of the LM and K-test. We consider the following sequence of local alternatives:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_{1T} : \theta = \theta^{(T)} \equiv \theta_0 + \frac{\gamma}{\delta_T}$$

where γ is a fixed deterministic p-vector and δ_T a deterministic sequence such that $\delta_T \overset{T}{\to} \infty$. Our approach follows, for instance, Davidson (2000, chapter 12.4)[2]. Note that under the

[2]In this paper, we do not investigate the power properties of specification tests under a sequence of local misspecification alternatives, as done for instance in Newey (1985). Note again the distinction between the true value of the parameter (θ^0) and its value under the null (θ_0). See also the discussion in Hall (2005, 5.3) about the connection between a rejection of H_0 and a misspecified model. These considerations are beyond the scope of this paper.

alternative, for each T, the true value of the parameter $\theta^{(T)}$ depends on the sample size.

Let us recall first the definitions of the two test statistics, where $\hat{S}_T$ denotes a standard consistent estimator of the long-term covariance matrix $S(\theta_0)$:

Definition 2.1. *To test $H_0 : \theta = \theta_0$, (i) the LM statistic is defined as,*

$$LM(\theta_0) = T\overline{\phi}'_T(\theta_0)\hat{S}_T'^{-1/2}A_T(\theta_0)\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)$$

with $A_T(\theta_0) = \hat{S}_T^{-1/2}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}\hat{S}_T^{-1}\frac{\partial\overline{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\overline{\phi}'_T(\theta_0)}{\partial\theta}\hat{S}_T^{-1/2'}$.
(ii) and the K statistic is defined as,

$$K(\theta_0) = T\overline{\phi}'_T(\theta_0)\hat{S}_T'^{-1/2}\tilde{A}_T(\theta_0)\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)$$

with $\tilde{A}_T(\theta_0) = \hat{S}_T^{-1/2}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\left[\frac{\partial\tilde{\phi}'_T(\theta_0)}{\partial\theta}\hat{S}_T^{-1}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\tilde{\phi}'_T(\theta_0)}{\partial\theta}\hat{S}_T^{-1/2'}$.

The above definitions emphasize that the only difference between the two test statistics is their weighting matrices, respectively $A_T(\theta_0)$ and $\tilde{A}_T(\theta_0)$ for LM and K. The main result of this section, theorem 2.5, compares the powers of the above test statistics against sequences of local alternatives (when varying δ_T and γ, respectively the rate and the direction of local departure). To precisely understand when and how the LM and K behave differently, we study first the asymptotic behavior of the key elements defining them: $\left[\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)\right]$, the corrected Jacobian and the matrices $A_T(\theta_0)$ and $\tilde{A}_T(\theta_0)$. The following theorems collect these results.

Theorem 2.1. *(Asymptotic behavior of $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)$)*
Under H_{1T}:
(i) If $\delta_T = \sqrt{T}$,

$$\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) \sim \mathcal{N}(0, I) + S^{-1/2}(\theta_0)c \quad \textit{where } c' = \left[\gamma'\frac{\partial\rho'_1(\theta_0)}{\partial\theta}\ 0\right]$$

(ii) If $\delta_T = \lambda_T$,
- If $\gamma \in Im\left[\frac{\partial\rho_1(\theta_0)}{\partial\theta'}\right]$, $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) = \mathcal{O}_P\left(\frac{\sqrt{T}}{\lambda_T}\right)$
- If $\gamma \in Im\left[\frac{\partial\rho_1(\theta_0)}{\partial\theta'}\right]^{\perp}$, $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) \sim \mathcal{N}(0, I) - S^{-1/2}(\theta_0)c$ where $c' = \left[0\ \gamma'\frac{\partial\rho'_2(\theta_0)}{\partial\theta}\right]$

Next, we show that the corrected Jacobian does not behave in a standard way with nearly-weak identification.

Theorem 2.2. *(Asymptotic behavior of the corrected Jacobian)*
Under H_{1T}*:*
(i) If $\delta_T = \sqrt{T}$*:*

$$\sqrt{T}\frac{\partial \tilde{\phi}_T(\theta_0)}{\partial \theta'} \sim \begin{pmatrix} \sqrt{T}\frac{\partial \rho_1(\theta_0)}{\partial \theta'} \\ \lambda_T \frac{\partial \rho_2(\theta_0)}{\partial \theta'} \end{pmatrix}$$

(ii) If $\delta_T = \lambda_T$ *and* $\lambda_T^2 >> \sqrt{T}$ *(nearly-strong identification):*

$$\sqrt{T}\frac{\partial \tilde{\phi}_T(\theta_0)}{\partial \theta'} \sim \begin{pmatrix} \sqrt{T}\frac{\partial \rho_1(\theta_0)}{\partial \theta'} \\ \lambda_T \frac{\partial \rho_2(\theta_0)}{\partial \theta'} \end{pmatrix}$$

(iii) If $\delta_T = \lambda_T$ *and* $\lambda_T^2 << \sqrt{T}$ *(nearly-weak identification):*

$$\sqrt{T}\frac{\partial \tilde{\phi}_T(\theta_0)}{\partial \theta'} \sim \begin{pmatrix} \sqrt{T}\frac{\partial \rho_1(\theta_0)}{\partial \theta'} \\ \frac{\sqrt{T}}{\lambda_T} B(\theta_0, \gamma) \end{pmatrix}$$

where $B(\theta_0, \gamma)$ *is the* $(k_2 \times p)$*-matrix with* j^{th} *column defined as*
$B_j = Cov\left(\sqrt{T}\frac{\partial \bar{\phi}_{2T}(\theta_0)}{\partial \theta_j}, \sqrt{T}\bar{\phi}_T(\theta_0)\right) S^{-1}(\theta_0) \begin{pmatrix} \frac{\partial \rho_1(\theta_0)}{\partial \theta'}\gamma \\ 0 \end{pmatrix}$ *for* $j = 1, \cdots, p$.

Theorem 2.3. *(Asymptotic behavior of the matrix* $A_T(\theta_0)$*)*
$A_T(\theta_0)$ *is asymptotically equivalent to the following projection (full-column rank) matrix*

$$A(\theta_0) = S^{-1/2}(\theta_0)\frac{\partial \rho(\theta_0)}{\partial \theta'}\left[\frac{\partial \rho'(\theta_0)}{\partial \theta} S^{-1}(\theta_0)\frac{\partial \rho(\theta_0)}{\partial \theta'}\right]^{-1}\frac{\partial \rho'(\theta_0)}{\partial \theta} S^{-1/2\prime}(\theta_0)$$

Next, we show that the weighting matrix $\tilde{A}_T(\theta_0)$ inherits the non-standard behavior of the corrected Jacobian with nearly-weak identification, while it behaves as $A_T(\theta_0)$ with (nearly)-strong identification.

Theorem 2.4. *(Asymptotic behavior of the matrix* $\tilde{A}_T(\theta_0)$*)*
(i) If $\delta_T = \sqrt{T}$*,* $\tilde{A}_T(\theta_0)$ *is asymptotically equivalent to the projection matrix* $A(\theta_0)$*.*
(ii) If $\delta_T = \lambda_T$ *and* $\lambda_T^2 >> \sqrt{T}$*,* $\tilde{A}_T(\theta_0)$ *is asymptotically equivalent to the projection matrix* $A(\theta_0)$*.*

(iii) If $\delta_T = \lambda_T$ *and* $\lambda_T^2 \leq\leq \sqrt{T}$*:*

- when N *is full-column rank,* $\tilde{A}_T(\theta_0)$ *is asymptotically equivalent to*

$$\tilde{A}_T(\theta_0) \sim S^{-1/2}(\theta_0)N\left[N'S^{-1}(\theta_0)N\right]^{-1}N'S^{-1/2\prime}(\theta_0)$$

$$\text{with } N \equiv \begin{pmatrix} \frac{\partial \rho_1(\theta_0)}{\partial \theta'} \\ B(\theta_0, \gamma) \end{pmatrix} \text{ and } B(\theta_0, \gamma) \text{ has been defined in theorem 2.2}$$

- when N *is not full-column rank,* $\tilde{A}_T(\theta_0)$ *is not asymptotically equivalent to a projection matrix of rank* p*.*

The question of interest is to determine when the asymptotic equivalent matrix of $\tilde{A}_T(\theta_0)$ is a projection matrix of rank p and when it is not. In general, we cannot answer this question. There is at least one case where we can conclude that $\tilde{A}_T(\theta_0)$ is not asymptotically equivalent to a projection matrix. This happens when γ is not spanned by the column-space[3] defined by $[\partial \rho_1'(\theta_0)/\partial\theta]$, that is when γ is not identified by the standard group of moment conditions. More formally,

$$\begin{aligned} \gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right]^{\perp} &\Longrightarrow \frac{\partial \rho_1(\theta_0)}{\partial \theta'}\gamma = 0 \\ &\Longrightarrow N = \begin{bmatrix} \frac{\partial \rho_1(\theta_0)}{\partial \theta'} \\ 0 \end{bmatrix} \text{ with } Rank\ N < p \\ &\Longrightarrow \left[N'S^{-1}(\theta_0)N\right] \text{ is not invertible} \\ &\Longrightarrow \tilde{A}_T(\theta_0) \text{ is not asymptotically full-column rank, hence not} \\ &\qquad \text{equivalent to a projection matrix of rank } p. \end{aligned}$$

We now state the main result of this chapter.

Theorem 2.5. *(Power of* LM *and* K *test statistics)*
(i) With strong identification (only standard moment conditions), $LM(\theta_0)$ *and* $K(\theta_0)$ *are asymptotically equivalent. They have the following power against local alternatives* H_{1T} *at rate* $\delta_T = \sqrt{T}$*:*

$$K(\theta_0) \sim LM(\theta_0) \xrightarrow{d} \chi_p^2(\mu) \qquad \textit{(under } H_{1T}\textit{)}$$

[3]Recall from Chapter 1 that the column-space defined by a matrix M is denoted as $Im[M]$

$$\textit{with} \quad \mu = \gamma' \frac{\partial \rho'(\theta_0)}{\partial \theta} S^{-1}(\theta_0) \frac{\partial \rho(\theta_0)}{\partial \theta'} \gamma \quad \textit{and} \quad \frac{\partial \rho(\theta_0)}{\partial \theta'} \gamma \neq 0 \ \ \forall \gamma$$

(ii) With nearly-strong identification ($\lambda_T^2 >> \sqrt{T}$), $LM(\theta_0)$ and $K(\theta_0)$ are asymptotically equivalent. They have the following power against local alternatives H_{1T}:

- when $\delta_T = \sqrt{T}$ and $\gamma \in Im\,[\partial \rho_1'(\theta_0)/\partial \theta]$:

$$K(\theta_0) \sim LM(\theta_0) \xrightarrow{d} \chi_p^2(\mu) \qquad \textit{(under } H_{1T}\textit{)}$$

$$\textit{with} \quad \mu = [\gamma' \frac{\partial \rho_1'(\theta_0)}{\partial \theta} \ 0] S^{-1}(\theta_0) \begin{bmatrix} \partial \rho_1(\theta_0)/\partial \theta' \gamma \\ 0 \end{bmatrix} \quad \textit{and} \quad \frac{\partial \rho_1(\theta_0)}{\partial \theta'} \gamma \neq 0$$

- when $\delta_T = \sqrt{T}$ and $\gamma \in Im\,[\partial \rho_1'(\theta_0)/\partial \theta]^{\perp}$, the power is equal to the size:

$$K(\theta_0) \sim LM(\theta_0) \xrightarrow{d} \chi_p^2 \qquad \textit{(under } H_{1T}\textit{)}$$

- when $\delta_T = \lambda_T$ and $\gamma \in Im\,[\partial \rho_1'(\theta_0)/\partial \theta]$:

$$K(\theta_0) \sim LM(\theta_0) \sim \mathcal{O}_P\left(\frac{T}{\lambda_T^2}\right) \quad \textit{and} \quad K(\theta_0) > 0 \qquad \textit{(under } H_{1T}\textit{)}$$

- when $\delta_T = \lambda_T$ and $\gamma \in Im\,[\partial \rho_1'(\theta_0)/\partial \theta]^{\perp}$:

$$K(\theta_0) \sim LM(\theta_0) \xrightarrow{d} \chi_p^2(\mu) \qquad \textit{(under } H_{1T}\textit{)}$$

$$\textit{with} \quad \mu = [0 \ \gamma' \frac{\partial \rho_2'(\theta_0)}{\partial \theta}] S^{-1}(\theta_0) \begin{bmatrix} 0 \\ \partial \rho_2(\theta_0)/\partial \theta' \gamma \end{bmatrix} \quad \textit{and} \quad \frac{\partial \rho_2(\theta_0)}{\partial \theta'} \gamma \neq 0$$

(iii) With nearly-weak identification ($\lambda_T^2 \leq\leq \sqrt{T}$), $LM(\theta_0)$ and $K(\theta_0)$ are not asymptotically equivalent. $LM(\theta_0)$ has the same asymptotic behavior and power as in case (ii).

- when $\delta_T = \sqrt{T}$: same as the similar case in (ii)

- when $\delta_T = \lambda_T$ and $\gamma \in Im\,[\partial \rho_1'(\theta_0)/\partial \theta]$: we cannot conclude about the asymptotic behavior of $\tilde{A}_T(\theta_0)$

- when $\delta_T = \lambda_T$ and $\gamma \in Im\,[\partial \rho_1'(\theta_0)/\partial \theta]^{\perp}$: $\tilde{A}_T(\theta_0)$ is not asymptotically full-column rank, hence not asymptotically equivalent to a projection matrix of rank p.

With strong identification, both tests have power against every direction of the local alternatives at the standard rate $\sqrt{T}$. When strong identification fails, this is not the case anymore. In particular, the framework of chapter 1 allows us to find standard directions (at rate $\sqrt{T}$)

against which the tests have only power equal to the size. We can also see that the tests have some power against slower alternatives (at rate λ_T): power that may depend (again) on the direction of departure from the null hypothesis. Such a power study is only possible because we decided to go one step further in the specification of the identification issues.

3 Testing hypotheses on subvectors

So far, we have focused on testing jointly the entire vector of the structural parameters θ. We might also be interested in testing a subset of these parameters, say $H_0^* : \beta = \beta_0$ when $\theta = (\alpha' \ \beta')'$. To do so, Kleibergen (2005) needs an additional assumption ensuring the full rank of the partial expected Jacobian with respect to the free parameters:

Assumption 4. *(Full rank of the partial expected Jacobian from Kleibergen (2005))*

$$\lim_{T\to\infty} E\left[\frac{\partial \overline{\phi}_T(\theta)}{\partial \alpha'}\right] \quad \textit{is a continuous function of } \theta \textit{ and has full rank at } \theta_0 = (\alpha_0' \ \beta_0')'.$$

Checking the validity of the above assumption involves several difficulties. First, of course, θ_0 is partially unknown under H_0^*. But, more generally, as mentioned p1111 in Kleibergen (2005), "it is not always straightforward to determine the parameters for which the assumption is satisfied".

However, in the framework of chapter 1, we do not meet such difficulties. After the convenient reparametrization, assumption 4 naturally holds for each (new) subvector identified by a group of moment conditions. Post-multiplying the initial Jacobian matrix by the matrix of reparametrization allows us to reinterpret the new Jacobian matrix as in assumption 4:

$$E\left[\frac{\partial \overline{\phi}_T(\theta)}{\partial \theta'}\right] R^0 = E\left[\frac{\partial \overline{\phi}_T(\theta)}{\partial \eta_1'} \ \frac{\partial \overline{\phi}_T(\theta)}{\partial \eta_2'}\right] \equiv \left[\frac{\partial \rho(\theta)}{\partial \eta_1'} \ \frac{\partial \rho(\theta)}{\partial \eta_2'}\right]$$

and each submatrix $[\partial \rho(\theta_0)/\partial \eta_i']$ has full column rank. In other words, testing the entire subvector η_1 or the entire subvector η_2 with the standard LM procedure works without any additional hypothesis.

Finally, note that, in general, an additional assumption is required when testing linear combinations of the structural parameters. This is to avoid perverse asymptotic correlations hap-

pening because of the multiplicity of rates of convergence. See also section 4 in chapter 1 for further details on Wald testing any transformation of the parameters.

4 Conclusion

In this chapter, we have performed a comparative power study between the standard GMM-LM test and its correction proposed by Kleibergen (2005).

We have shown that this correction does have asymptotic consequences, especially with heterogeneous identification patterns. Hence, we recommend carefulness, especially when instruments of heterogeneous quality are used. Moreover, we also recommend using the framework of chapter 1. As shown in this chapter, it is not much more involved in terms of specifying the identification issues. In addition, not only it enables the use of (valid) standard test procedures (like GMM-LM and Wald), but also it helps identify the directions against which the tests have power.

In terms of testing hypothesis on subvectors, the superiority of the framework of chapter 1 is clear. The reparametrization (see section 3 in chapter 1) precisely identifies the directions in the parameter space for which the standard GMM-LM test can be performed. In particular, no additional assumption on the free (remaining) parameters is required as for the K-test of Kleibergen (2005). More generally, this framework also deals with (nonlinear) transformations of the structural parameters. This is beyond the scope of Kleibergen (2005) (see section 4 in chapter 1).

Appendix

Proof of Theorem 2.1 *(Behavior of $\sqrt{T}\hat{S}_T^{-1/2}(\theta_0)\overline{\phi}_T(\theta_0)$)*:
The application of the mean-value theorem gives:

$$\overline{\phi}_T(\theta_0) = \overline{\phi}_T(\theta^{(T)} - \gamma/\delta_T) = \overline{\phi}_T(\theta^{(T)}) + \frac{\partial\overline{\phi}_T^*}{\partial\theta'}(\theta_0 - \theta^{(T)}) = \overline{\phi}_T(\theta^{(T)}) - \frac{\partial\overline{\phi}_T^*}{\partial\theta'}\frac{\gamma}{\delta_T}$$

with $\left[\frac{\partial\overline{\phi}_T^*}{\partial\theta'}\right]$ the Jacobian matrix evaluated at a vector with each component between θ_0 and $\theta^{(T)}$. In addition, we have:

$$\begin{aligned} Var[\sqrt{T}\overline{\phi}_T(\theta_0)] &= Var\left(\sqrt{T}\overline{\phi}(\theta^{(T)}) - \sqrt{T}\frac{\partial\overline{\phi}_T^*}{\partial\theta'}\frac{\gamma}{\delta_T}\right) \\ &= Var[\sqrt{T}\overline{\phi}_T(\theta^{(T)})] + Var\left(\sqrt{T}\frac{\partial\overline{\phi}_T^*}{\partial\theta'}\frac{\gamma}{\delta_T}\right) \\ &\quad -2Cov\left(\sqrt{T}\overline{\phi}_T(\theta^{(T)}), \sqrt{T}\frac{\partial\overline{\phi}_T^*}{\partial\theta'}\frac{\gamma}{\delta_T}\right) \\ &\sim Var[\sqrt{T}\overline{\phi}_T(\theta^{(T)})] \end{aligned}$$

Finally,

$$\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) = \sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta^{(T)}) - \sqrt{T}\hat{S}_T^{-1/2}\frac{\partial\overline{\phi}_T^*}{\partial\theta'}\frac{\gamma}{\delta_T}$$

We can also deduce that under H_{1T}[4]:

$$RHS(1) \sim \mathcal{N}(0, I) \quad \text{and} \quad RHS(2) \sim S^{-1/2}(\theta_0)\begin{pmatrix} \frac{\sqrt{T}}{\delta_T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'}\gamma \\ \frac{\lambda_T}{\delta_T}\frac{\partial\rho_2(\theta_0)}{\partial\theta'}\gamma \end{pmatrix}$$

(i) $\delta_T = \sqrt{T}$:

$$\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) \sim \mathcal{N}(0, I) - S^{-1/2}(\theta_0)\begin{pmatrix} \frac{\partial\rho_1(\theta_0)}{\partial\theta'}\gamma \\ 0 \end{pmatrix}$$

(ii) $\delta_T = \lambda_T$:

$$RHS(2) \sim S^{-1/2}(\theta_0)\begin{pmatrix} \frac{\sqrt{T}}{\lambda_T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'}\gamma \\ \frac{\partial\rho_2(\theta_0)}{\partial\theta'}\gamma \end{pmatrix}$$

[4]Note that the result for $RHS(1)$ is a little more involved because we now deal with an element of a triangular array. See Davidson (2000) p298 for a similar discussion and appropriate regularity conditions.

- If $\gamma \in Im\left[\frac{\partial \rho_1(\theta_0)}{\partial \theta'}\right]$, $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) = \mathcal{O}_p(1) + \mathcal{O}_p\left(\frac{\sqrt{T}}{\lambda_T}\right) = \mathcal{O}_p\left(\frac{\sqrt{T}}{\lambda_T}\right)$

- If $\gamma \in Im\left[\frac{\partial \rho_1(\theta_0)}{\partial \theta'}\right]^{\perp}$, $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) \sim \mathcal{N}(0, I) - S^{-1/2}(\theta_0)\begin{pmatrix} 0 \\ \frac{\partial \rho_2(\theta_0)}{\partial \theta'}\gamma \end{pmatrix}$ ■

Proof of Theorem 2.2 *(Behavior of the corrected Jacobian)*:
At the beginning of this proof, we treat each component of the moment conditions separately: therefore, the index $i = 1, \cdots, K$ refers to the component and not to the group of moment conditions as in the main text. Recall first the definition of the corrected Jacobian:

$$\sqrt{T}\frac{\partial \tilde{\phi}_{iT}(\theta_0)}{\partial \theta_j} = \sqrt{T}\frac{\partial \overline{\phi}_{iT}(\theta_0)}{\partial \theta_j} - B_{ij}\sqrt{T}\overline{\phi}_T(\theta_0)$$

where $B_{ij} = Cov\left(\sqrt{T}\frac{\partial \overline{\phi}_{iT}(\theta_0)}{\partial \theta_j}, \sqrt{T}\overline{\phi}_T(\theta_0)\right) S^{-1}(\theta_0)$ for $i = 1, \cdots, K$ and $j = 1, \cdots, p$.
The application of the mean-value theorem gives:

$$\frac{\partial \overline{\phi}_{iT}(\theta_0)}{\partial \theta_j} = \frac{\partial \overline{\phi}_{iT}(\theta^{(T)})}{\partial \theta_j} - \frac{\partial}{\partial \theta'}\left[\frac{\partial \overline{\phi}_{iT}}{\partial \theta_j}\right]^{**}\frac{\gamma}{\delta_T}$$

where $[.]^{**}$ denotes the Hessian evaluated at a vector whose components are between θ_0 and $\theta^{(T)}$. Recall also (see proof of theorem 2.1):

$$\sqrt{T}\overline{\phi}_T(\theta_0) = \sqrt{T}\overline{\phi}_T(\theta^{(T)}) - \frac{\partial \overline{\phi}_T^*}{\partial \theta'}\sqrt{T}\frac{\gamma}{\delta_T}$$

We deduce:

$$\sqrt{T}\frac{\partial \tilde{\phi}_{iT}(\theta_0)}{\partial \theta_j} = \sqrt{T}\frac{\partial \overline{\phi}_{iT}(\theta^{(T)})}{\partial \theta_j} - \frac{\partial}{\partial \theta'}\left[\frac{\partial \overline{\phi}_{iT}}{\partial \theta_j}\right]^{**}\sqrt{T}\frac{\gamma}{\delta_T} - B_{ij}\sqrt{T}\overline{\phi}_T(\theta^{(T)}) + B_{ij}\frac{\partial \overline{\phi}_T^*}{\partial \theta'}\sqrt{T}\frac{\gamma}{\delta_T}$$

Define first the block-diagonal matrix Λ_T (as in the proof of lemma A in chapter 1),
$\Lambda_T = \begin{pmatrix} Id_{k_1} & 0 \\ 0 & \frac{\lambda_T}{\sqrt{T}}Id_{k_2} \end{pmatrix}$ and μ_{iT} the rate of convergence associated to the i^{th} component,
$\mu_{iT} = \sqrt{T}$ for $1 \leq i \leq k_1$ and $\mu_{iT} = \lambda_T$ for $k_1 + 1 \leq i \leq K$.
- with assumption 2(iv) from chapter 1 (about the Plim of the well-scaled Jacobian) we get: $RHS(1) \sim \mu_{iT}\frac{\partial \rho_i(\theta^{(T)})}{\partial \theta_j}$.
- with assumption 3(ii) from chapter 1 (about the Plim of the well-scaled Hessian) we get: $RHS(2) \sim \mu_{iT}H_i\frac{\gamma}{\delta_T}$ for some fixed matrix H_i. This is dominated by $RHS(1)$.
- $RHS(3) \sim B_{ij}\sqrt{T}\Lambda_T\rho(\theta^{(T)})$ and $\rho(\theta^{(T)}) = 0$ under H_{1T}.
- $RHS(4) \sim B_{ij}\Lambda_T\frac{\partial \rho(\theta^{(T)})}{\partial \theta'}\frac{\gamma}{\delta_T}$.

Finally,

$$\sqrt{T}\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} \sim \mu_{iT}\frac{\partial\rho_i(\theta^{(T)})}{\partial\theta_j} + B_{ij}\begin{bmatrix} \sqrt{T}\frac{\partial\rho_1(\theta^{(T)})}{\partial\theta'} \\ \lambda_T\frac{\partial\rho_2(\theta^{(T)})}{\partial\theta'} \end{bmatrix}\frac{\gamma}{\delta_T}$$

• Study of the terms of the RHS:

- when $\mu_{iT} = \sqrt{T}$ (ie the i-th component is strong):

$$\sqrt{T}\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} = \sqrt{T}\frac{\partial\rho_i(\theta^{(T)})}{\partial\theta_j} + B_{ij}\begin{bmatrix} \sqrt{T}\frac{\partial\rho_1(\theta^{(T)})}{\partial\theta'} \\ \lambda_T\frac{\partial\rho_2(\theta^{(T)})}{\partial\theta'} \end{bmatrix}\frac{\gamma}{\delta_T} \Rightarrow \frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} \sim \frac{\partial\rho_i(\theta_0)}{\partial\theta_j}$$

- when $\mu_{iT} = \lambda_T$ (ie the i-th component is nearly-weak):

$$\sqrt{T}\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} \sim \lambda_T\frac{\partial\rho_i(\theta_0)}{\partial\theta_j} + B_{ij}\begin{bmatrix} \sqrt{T}\frac{\partial\rho_1(\theta^{(T)})}{\partial\theta'} \\ \lambda_T\frac{\partial\rho_2(\theta^{(T)})}{\partial\theta'} \end{bmatrix}\frac{\gamma}{\delta_T}$$

* if $\delta_T = \sqrt{T}$:

$$\sqrt{T}\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} \sim \lambda_T\frac{\partial\rho_i(\theta_0)}{\partial\theta_j}$$

* if $\delta_T = \lambda_T$ and $\lambda_T >> \sqrt{T}/\lambda_T$:

$$\sqrt{T}\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} \sim \lambda_T\frac{\partial\rho_i(\theta_0)}{\partial\theta_j}$$

* if $\delta_T = \lambda_T$ and $\lambda_T << \sqrt{T}/\lambda_T$:

$$\sqrt{T}\frac{\partial\tilde{\phi}_{iT}(\theta_0)}{\partial\theta_j} \sim B_{ij}\begin{bmatrix} \frac{\sqrt{T}}{\lambda_T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'} \\ \frac{\partial\rho_2(\theta_0)}{\partial\theta'} \end{bmatrix}\gamma$$

• Extension to treat all components simultaneously: we are back to our regular formalism where the indexes 1 and 2 refer to the groups of moment conditions.
The above calculations lead to:

$$\sqrt{T}\frac{\partial\tilde{\phi}_T}{\partial\theta'} \sim \begin{pmatrix} \sqrt{T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'} \\ \lambda_T\frac{\partial\rho_2(\theta_0)}{\partial\theta'} + B_T(\theta_0,\gamma,\delta_T) \end{pmatrix}$$

where $B_T(.)$ is the $(k_2 \times p)$-matrix with j^{th} column defined as

$$B_{Tj} = Cov(\sqrt{T}\frac{\partial\bar{\phi}_{2T}(\theta_0)}{\partial\theta_i}, \sqrt{T}\bar{\phi}_T(\theta_0))S^{-1}(\theta_0)\begin{pmatrix} \frac{\sqrt{T}}{\delta_T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'}\gamma \\ \frac{\lambda_T}{\delta_T}\frac{\partial\rho_2(\theta_0)}{\partial\theta'}\gamma \end{pmatrix} \text{ for } j = 1, \cdots, p.$$

To conclude:

(i) $\delta_T = \sqrt{T}$ or (ii) $\delta_T = \lambda_T$ and $\lambda_T^2 >> \sqrt{T}$:

$$\sqrt{T}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'} \sim \begin{pmatrix} \sqrt{T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'} \\ \lambda_T\frac{\partial\rho_2(\theta_0)}{\partial\theta'} \end{pmatrix}$$

(iii) $\delta_T = \lambda_T$ and $\lambda_T^2 << \sqrt{T}$:

$$\sqrt{T}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'} \sim \begin{pmatrix} \sqrt{T}\frac{\partial\rho_1(\theta_0)}{\partial\theta'} \\ \frac{\sqrt{T}}{\lambda_T}B(\theta_0, \gamma) \end{pmatrix}$$

■

Proof of Theorem 2.3 *(Asymptotic behavior of the matrix $A_T(\theta_0)$)*:

Recall the mean-value theorem on the Jacobian:

$$\sqrt{T}\frac{\partial\bar{\phi}_{iT}(\theta_0)}{\partial\theta_j} = \sqrt{T}\frac{\partial\bar{\phi}_T(\theta^{(T)})}{\partial\theta_j} - \sqrt{T}\frac{\partial}{\partial\theta'}\left[\frac{\partial\bar{\phi}_{iT}}{\partial\theta_j}\right]^{**}\frac{\gamma}{\delta_T} \sim \mu_{iT}\frac{\partial\rho_i(\theta^{(T)})}{\partial\theta_j}$$

Λ_T, as defined in the proof of theorem 2.2, is invertible for any sample size T. We deduce:

$$\begin{aligned} A_T(\theta_0) &= \hat{S}_T^{-1/2}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\Lambda_T\left[\Lambda_T\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta}\hat{S}_T^{-1}\frac{\partial\bar{\phi}_T(\theta_0)}{\partial\theta'}\Lambda_T\right]^{-1}\Lambda_T\frac{\partial\bar{\phi}_T'(\theta_0)}{\partial\theta}\hat{S}_T^{-1/2\prime} \\ &\sim S^{-1/2}(\theta_0)\frac{\partial\rho(\theta_0)}{\partial\theta'}\left[\frac{\partial\rho'(\theta_0)}{\partial\theta}S^{-1}(\theta_0)\frac{\partial\rho(\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial\rho'(\theta_0)}{\partial\theta}S^{-1/2\prime}(\theta_0) \end{aligned}$$

which is a projection matrix because by assumption $[\partial\rho(\theta_0)/\partial\theta']$ is full column-rank. ■

Proof of Theorem 2.4 *(Asymptotic behavior of matrix $\tilde{A}_T(\theta_0)$)*:
(i) $\delta_T = \sqrt{T}$:

$$\begin{aligned}
\tilde{A}_T(\theta_0) &= \hat{S}_T^{-1/2}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\left(\frac{\partial\tilde{\phi}_T'(\theta_0)}{\partial\theta}\hat{S}_T^{-1}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\right)^{-1}\frac{\partial\tilde{\phi}_T'(\theta_0)}{\partial\theta}\hat{S}_T^{-1/2'} \\
&= \hat{S}_T^{-1/2}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\Lambda_T\left(\Lambda_T\frac{\partial\tilde{\phi}_T'(\theta_0)}{\partial\theta}\hat{S}_T^{-1}\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\Lambda_T\right)^{-1}\Lambda_T\frac{\partial\tilde{\phi}_T'(\theta_0)}{\partial\theta}\hat{S}_T^{-1/2'} \\
&\sim S^{-1/2}(\theta_0)\frac{\partial\rho(\theta_0)}{\partial\theta'}\left(\frac{\partial\rho'(\theta_0)}{\partial\theta}S^{-1}(\theta_0)\frac{\partial\rho(\theta_0)}{\partial\theta'}\right)^{-1}\frac{\partial\rho'(\theta_0)}{\partial\theta}S^{-1/2'}(\theta_0)
\end{aligned}$$

where Λ_T is the invertible matrix defined in the proof of theorem 2.3. The last matrix is $A(\theta_0)$, a projection matrix of rank p because $S^{-1/2}(\theta_0)\partial\rho/\partial\theta'$ is full-column rank by assumption.

(ii) $\delta_T = \lambda_T$ and $\lambda_T^2 >> \sqrt{T}$: similar to (i).

(iii) $\delta_T = \lambda_T$ and $\lambda_T^2 << \sqrt{T}$, we proceed as in the case (i). From theorem 2.2(iii), we have:

$$\frac{\partial\tilde{\phi}_T(\theta_0)}{\partial\theta'}\begin{bmatrix} Id_{s1} & 0 \\ 0 & \lambda_T Id_{p-s1}\end{bmatrix} \sim \begin{pmatrix} \frac{\partial\rho_1(\theta_0)}{\partial\theta'} \\ B(\theta_0,\gamma)\end{pmatrix} \equiv N$$

- When N is full-column rank, $[N'S^{-1}(\theta_0)N]$ is full rank, hence invertible. We have:

$$\tilde{A}_T(\theta_0) \sim S^{-1/2}(\theta_0)N\left[N'S^{-1}(\theta_0)N\right]^{-1}N'S^{-1/2'}(\theta_0)$$

- When N is not full-column rank, $[N'S^{-1}(\theta_0)N]$ is not invertible. Then, $\tilde{A}_T(\theta_0)$ is not asymptotically equivalent to a projection matrix. ■

Proof of Theorem 2.5 *(Power of LM and K test statistics)*:
(i) From theorems 2.3 and 2.4(i): $A_T(\theta_0) \sim \tilde{A}_T(\theta_0) \sim A(\theta_0)$

$$\Rightarrow LM(\theta_0) \sim K(\theta_0) \sim T\overline{\phi}_T'(\theta_0)\hat{S}_T^{-1/2\prime}A(\theta_0)\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)$$

From theorem 2.1(i): $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) \sim \mathcal{N}(0,I) - S^{-1/2}(\theta_0)c$ where $c' = \left[\gamma' \frac{\partial \rho'(\theta_0)}{\partial \theta}\right]$. We get (after applying Corollary B.3 from Gouriéroux and Monfort (1995)):

$$\mu = c'S^{-1/2\prime}(\theta_0)A(\theta_0)S^{-1/2}(\theta_0)c = \frac{1}{2}\gamma'\frac{\partial \rho'(\theta_0)}{\partial \theta}S^{-1}(\theta_0)\frac{\partial \rho(\theta_0)}{\partial \theta'}\gamma$$

(ii) From theorems 2.3 and 2.4(i) and (ii): $A(\theta_0) \sim \tilde{A}_T(\theta_0) \sim A(\theta_0) \Rightarrow K(\theta_0) \sim LM(\theta_0)$.
- if $\delta_T = \sqrt{T}$, from theorem 2.1(i): $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) \sim \mathcal{N}(0,I) - S^{-1/2}(\theta_0)c$ where $c' = \left[\gamma' \frac{\partial \rho_1'(\theta_0)}{\partial \theta} \; 0\right]$. Then the calculation is similar to (i).
Note also that: $\gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right] \Rightarrow \frac{\partial \rho_1(\theta_0)}{\partial \theta'}\gamma \neq 0$; and $\gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right]^{\perp} \Rightarrow \frac{\partial \rho_1(\theta_0)}{\partial \theta'}\gamma = 0$.
- if $\delta_T = \lambda_T$ and $\gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right]$, from theorem 2.1(ii):
$\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) = \mathcal{O}_P\left(\frac{\sqrt{T}}{\lambda_T}\right) \Rightarrow LM(\theta_0) = \mathcal{O}_P\left(\frac{T}{\lambda_T^2}\right)$.
Also, $LM(\theta_0) > 0$ as a quadratic form with a positive definite weighting matrix $A(\theta_0)$
- if $\delta_T = \lambda_T$ and $\gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right]^{\perp}$, from theorem 2.1(ii):
$\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0) = \mathcal{N}(0,I) - S^{-1/2}(\theta_0)c$ where $c' = \left[0 \; \gamma' \frac{\partial \rho_2'(\theta_0)}{\partial \theta}\right]$. Then the calculation is similar to (i).

(iii) - From theorems 2.1 and 2.3, there is no distinction in the asymptotic behavior of $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)$ and $A_T(\theta_0)$ in cases where $\lambda_T^2 << \sqrt{T}$ or $\lambda_T^2 >> \sqrt{T}$. Hence, $LM(\theta_0)$ behaves similarly in cases (ii) and (iii).
- if $\delta_T = \sqrt{T}$, from theorems 2.1(i), 2.2(i) and 2.4(i), there is no distinction in the asymptotic behavior of $\sqrt{T}\hat{S}_T^{-1/2}\overline{\phi}_T(\theta_0)$ and $\tilde{A}_T(\theta_0)$ in cases where $\lambda_T^2 << \sqrt{T}$ or $\lambda_T^2 >> \sqrt{T}$. Hence, $K(\theta_0)$ behaves similarly as in the similar case in (ii).
- if $\delta_T = \lambda_T$ and $\gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right]$, from theorem 2.4(iii) the asymptotic behavior of $\tilde{A}_T(\theta_0)$ is not clear. So we cannot conclude.
- if $\delta_T = \lambda_T$ and $\gamma \in Im\left[\frac{\partial \rho_1'(\theta_0)}{\partial \theta}\right]^{\perp}$, from theorem 2.4(iii) and the comment following it, $\tilde{A}_T(\theta_0)$ is not asymptotically full-column rank, hence not asymptotically equivalent to a projection matrix of rank p. ■

Chapitre III

Inference on the Parameter Ratio with Applications to Weak Identification

1 Introduction

Providing reliable inference about the parameter of interest has always been a question of interest in econometrics. Confidence regions represent a convenient way to withdraw such information. For instance, these regions are built after inverting a Wald-type test statistic. It simply means that one collects all the values of the parameter for which the test is not significant. Under regularity conditions and smooth parameter functions, such confidence regions are bounded: intervals in the unidimensional case, and, more generally, ellipsoids.

Recently, interest has grown in providing (valid) inference when the identification of the parameter is not fully ensured. In this essay, we focus on the following ratio of parameters $(\mathcal{I}_1^{-1} \times \mathcal{I}_2)$, when the denominator $(\mathcal{I}_1)$ is *close* to singularity. We propose a new inference method, the Modified-Wald procedure. This method is based on the Wald statistic. The key idea consists in integrating the informational content of the null hypothesis of interest in the computation of its metric. This correction, while preserving the computational tractability of the method, allows for *unbounded* confidence regions when needed. Boundedness has become an issue since Dufour (1997). In the context of *local almost unidentification*, Dufour (1997) provides some results on the characterization of the confidence regions: under regularity conditions, these regions should be unbounded with non-zero probability. In particular, when identification fails, most Wald-type confidence sets have zero confidence level because they are almost surely bounded. By contrast, inference methods based on the Likelihood-ratio (LR) statistic do not encounter such problems.

One practical difficulty for applying LR statistics is that their calculations require both the unconstrained and the constrained estimators, whereas Wald-type statistics only involve the unconstrained one. This mainly explains the popularity of Wald-type inference methods, despite some well-known and documented drawbacks. Our Modified-Wald procedure is computationally friendly, just like the classic Wald method, and corrects for its lack of unboundedness. When identification fails at the frontier of the parameter space, in the spirit of Dufour (1997), we show that the probability of getting an unbounded confidence region reaches the upper bound of Dufour (1997). When identification issues are (artificially) tied to the sample size, in the spirit of the *Pitman drift*, this probability depends on the rate of convergence towards unidentification. For instance, with weak identification and a rate equal the square-root of the sample size, this probability is non-zero but smaller than the upper bound. Two applications help clarify these results: first, a simulation exercise is designed to

compare the inference properties of the Wald and Modified-Wald procedures with a bidimensional ratio, when identification fails at the frontier of the parameter space; second, the linear single-equation instrumental variables regression model is considered, when the identifying properties of the instruments may vary.

The remainder of this chapter is organized as follows: section 2 introduces our framework; in section 3, the Modified-Wald method is defined and its main properties are stated; in section 4, these methods and results are applied to two empirical econometric examples: a model-free ratio of parameter and a single-equation linear IV regression model; finally, section 5 concludes.

All the proofs and the figures are gathered in the appendix. We use the following notation throughout the chapter: $\otimes$ stands for the Kroneker product, $P_A = A(A'A)^{-1}A'$, and $M_A = I_n - P_A$ for a full column rank (n, r)-matrix A $(n \geq r)$ and the (n, n) identity matrix I_n. $\xrightarrow{p}$ indicates convergence in probability, and $\xrightarrow{d}$ convergence in distribution.

2 Framework

Consider the following vector of parameters, $\theta = vec[\beta_1 \ \beta_2]$[1] where β_1 is an invertible (r, r)-matrix and β_2 is a r-vector. We are interested in providing inference on the following ratio of parameters,

$$\underset{(r,1)}{\psi(\theta)} = \underset{(r,r)}{\beta_1^{-1}} \times \underset{(r,1)}{\beta_2} \tag{2.1}$$

defined for all $\theta \in \Theta$. Θ is an open subset of $\mathbb{R}^p$ such that $\Theta \subset \{\theta \in \mathbb{R}^p \ / \ det(\beta_1) \neq 0\}$, with $p = r(r+1)$.

The transformation $\psi(.)$ forms a set of r functionally independent constraints with $1 \leq r \leq p$, differentiable at all interior points of Θ; the Jacobian (r, p)-matrix $[\partial\psi(\theta)/\partial\theta']$ is assumed to have full rank r, at least in an open neighborhood of θ_0.

[1] The operator $vec(.)$ transforms a (m,n)-matrix into a (mn)-vector. The former is obtained after stacking all the columns of the matrix. For instance,

$$\text{if } M = \begin{pmatrix} 1 & 5 & 2 \\ 0 & 1 & 8 \end{pmatrix} \text{ then } vec(M) = [1\ 0\ 5\ 1\ 2\ 8]'$$

More specifically, we are interested in building confidence regions for the transformation $\psi(.)$ when ϑ_1 is (potentially) close to singularity (to be defined more precisely later). These regions are defined as the set of the values $\psi_0 \in \mathbb{R}^r$ for which the null hypothesis $H_0(\psi_0) : \psi(\theta) = \psi_0$ cannot be rejected at some chosen level of confidence.

Inference is drawn from n observations of the random variable X: it comes from a distribution P_θ on some measurable space $(\mathcal{X}, \mathcal{A})$ indexed by the parameter $\theta \in \Theta$. The data generating process (DGP) is represented by the point θ_0 which is assumed to be an inner point of the original parameter set.

Assumption 1. *Given the observed data $x = (x_1 \cdots x_n)'$ on the random variable X, $\hat{\theta}_n$ is an asymptotically normal estimator of θ_0 an interior point of Θ,*

$$\sqrt{n}\left[\hat{\theta}_n - \theta_0\right] \xrightarrow{d} \mathcal{N}(0, \Sigma_\theta)$$

where the asymptotic variance of $\hat{\theta}_n$, Σ_θ, is supposed to be known[2].

The asymptotic normality of the estimator $\hat{\theta}_n$ is the only assumption we require here. Commonly, confidence regions are built after inverting a Wald-type statistic[3]. The associated Wald-type confidence region is then defined as:

$$CR_W(\alpha) = \left\{ \psi_0 / n\left[\psi(\hat{\theta}_n) - \psi_0\right]' \Sigma_\psi^{-1}(\hat{\theta}_n)\left[\psi(\hat{\theta}_n) - \psi_0\right] \leq \chi_\alpha^2(r) \right\} \tag{2.2}$$

where $\chi_\alpha^2(r)$ denotes the $(1-\alpha)$-quantile of a Chi-square distribution with r degrees of freedom and $\Sigma_\psi(\hat{\theta}_n)$ is the estimated asymptotic variance of $\psi(\hat{\theta}_n)$ obtained by the delta-method:

$$\Sigma_\psi(\hat{\theta}_n) = \frac{\partial \psi(\hat{\theta}_n)}{\partial \theta'} \Sigma_\theta \frac{\partial \psi'(\hat{\theta}_n)}{\partial \theta} \tag{2.3}$$

The Wald statistic is natural and easy to implement since it does not involve the estimation of the constrained model, or estimation under the null hypothesis H_0. However, it (often)

[2] Note that the following theory is not affected if Σ_θ is unknown but can be consistently estimated.

[3] Note that the original Wald statistic (Wald (1949)) was defined for a parametric model $\mathcal{M} = \{f(x;\theta)|\theta \in \Theta\}$ and using the Maximum Likelihood estimator. Wald-type statistics use any consistent and asymptotically normal estimator of θ.

yields to ellipsoidal confidence regions which are symmetric and bounded. Boundedness has become a real issue since Dufour (1997), who shows the following necessary condition:

Theorem 2.1. *(Necessity for Unboundedness from Dufour (1997))*
When a locally almost unidentified parametric function has an unbounded range, under regularity conditions, any valid confidence set should have nonzero probability of being unbounded under any distribution compatible with the model.

A precise definition of *locally almost unidentified* is given in Dufour (1997). The idea is to consider series of parameters θ_n such that i) $\theta_n \in \Theta \; \forall \; n$; ii) $\psi(\theta_n) = \psi_0$; iii) θ_n converges to a (discontinuity) point at the frontier of the parameter space. In such a case, note that a valid confidence region CR_ψ with level $(1-\alpha)$ should be unbounded with probability as high as $(1-\alpha)$:

$$\lim_{n \to \infty} \inf P_{\theta_n} [\psi(\theta) \in CR_\psi] \geq 1 - \alpha$$

By contrast, the Wald-type confidence region (2.2) does not generally satisfy the above necessary condition. In general, $\Sigma_\psi(\theta_n)$ is a symmetric positive definite matrix and does not depend on ψ_0. In that case, $CR_W(\alpha)$ is a (bounded) ellipsoid. In particular, we can show that even though,

$$\forall \, \theta \in \Theta \quad P_{\theta_n} [\psi(\theta) \in CR_W(\alpha)] \overset{n \to \infty}{\longrightarrow} 1 - \alpha$$

we may have an invalid procedure in the sense that the infimum of its coverage probability is null,

$$\forall \, n \quad \inf_{\theta \in \Theta} P_{\theta_n} [\psi(\theta) \in CR_W(\alpha)] = 0 \tag{2.4}$$

The need for unboundedness (with nonzero probability) is fairly natural. Let consider the following unidimensional ratio, $\psi = \beta_2 / \beta_1$. The closer the true value of the denominator $(\beta_{1,0})$ is to zero, the less informative the estimator $(\hat{\beta}_{1n}, \hat{\beta}_{2n})$ is on the ratio ψ; hence the larger the associated confidence region should be. In the limit, when $(\beta_{1,0})$ is arbitrary close to zero, the confidence region should even become arbitrary large. Note that, inside the parameter space Θ, there exists some parameter settings with arbitrary small $|\beta_{1,0}|$: allowing for unbounded confidence regions is the only way to prevent (2.4) to happen. More intuitively, unbounded regions should be permitted for settings where the denominator β_1 is not significant.

3 The Modified-Wald procedure

3.1 Definition

Explaining the failure of Wald-(type) procedures is not obvious. However, one could intuitively think that Wald statistics do not fully incorporate the information available from the null hypothesis. We do not mean that constrained estimation should be considered here: this would actually kill the computational advantage of our procedure. This is rather related to what is known under the null hypothesis and what is not, or privileged directions in the parameter space. More precisely, the metric of the Wald statistic $\Sigma_{\psi}^{-1}(\theta)$ (see equation (2.3)) is fixed with respect to the null hypothesis of interest: it is a function of the entire parameter θ and not of the parameter of interest $\psi(\theta)$. Practically, if we were testing $H_0 : \psi(\theta) = \psi^*$, the metric would remain numerically the same for any ψ^*. In some sense, the metric of the Wald statistic is disconnected from the null hypothesis[4]. This motivates us to incorporate the information of the null hypothesis in the calculation of this metric.

In general, $\psi(\theta)$ does not constitute a complete reparametrization of the problem ($r \leq p$). So, one cannot directly map θ into some function ψ. Hence, we first need to *complete this partial* reparametrization. Define the new p-vector of parameters as,

$$\theta^* = \begin{pmatrix} \psi \\ vec(\beta_1) \end{pmatrix} \quad \text{with} \quad Rank\left[\frac{\partial \theta^*}{\partial \theta'}\right] = p \tag{3.1}$$

$\psi(.)$ is a r-vector representing the constrained directions (known under H_0) and the $(p-r)$-vector $vec(\beta_1)$ can be interpreted as the free directions (unknown under H_0). Basically, if we were working in a parametric model, the initial model $\mathcal{M} = \{f(x;\theta)|\theta \in \Theta\}$ is now replaced by $\mathcal{M} = \{f^*(x;\theta^*)|\theta^* \in \Theta\}$ which is obtained after a legitimate change of parameter. Since θ^* represents a legitimate reparametrization of the model, we now know that $\Sigma_{\psi}(\theta)$ can be reexpressed as a function of θ^* only.

[4]This observation is related to the work of Critchley, Marriott and Salmon (1996) in differential geometry. They note that the Wald statistic is not a genuine geometrical object: it is neither the squared length of a vector in a tangent space (since $\psi(\theta)$ does not belong to the tangent space), nor the squared distance between two points in a manifold (since it uses a fixed metric, whereas the metric should in general vary with ψ). For an introduction to these concepts, see Critchley *et al.* (1996) and for a more complete treatment see Amari (1990).

The estimator $\hat{\theta}_n^\star$ of the vector $\theta^\star$ naturally writes:

$$\hat{\theta}_n^\star = \begin{pmatrix} \hat{\psi}_n \\ vec(\hat{\beta}_{1n}) \end{pmatrix}$$

and its asymptotic variance is

$$\Sigma_{\theta^\star}(\hat{\theta}_n^\star) = \frac{\partial \hat{\theta}_n^\star}{\partial \theta'} \Sigma_\theta \frac{\partial \hat{\theta}_n^{\star\prime}}{\partial \theta} = \begin{pmatrix} \Sigma_\psi(\hat{\theta}_n^\star) & \Sigma_{12}(\hat{\theta}_n^\star) \\ \Sigma_{12}(\hat{\theta}_n^\star)' & \Sigma_{22}(\hat{\theta}_n^\star) \end{pmatrix} \tag{3.2}$$

$$\begin{aligned} \text{with} \quad \Sigma_\psi(\hat{\theta}_n^\star) &= \frac{\partial \psi(\hat{\theta}_n^\star)}{\partial \theta'} \Sigma_\theta \frac{\partial \psi'(\hat{\theta}_n^\star)}{\partial \theta} \\ \Sigma_{12}(\hat{\theta}_n^\star) &= \frac{\partial vec(\hat{\beta}_{1n})}{\partial \theta'} \Sigma_\theta \frac{\partial vec'(\hat{\beta}_{1n})}{\partial \theta} \\ \Sigma_{22}(\hat{\theta}_n^\star) &= \frac{\partial \psi(\hat{\theta}_n^\star)}{\partial \theta'} \Sigma_\theta \frac{\partial vec'(\hat{\beta}_{1n})}{\partial \theta} \end{aligned}$$

At this stage, it is important to stress that the completion of $\psi(\theta)$ is not unique. For instance, we could decide to impose the first r (constant) components evaluated at $\hat{\theta}_n$ to be orthogonal to the $(p-r)$ remaining ones with respect to the metric Σ_θ. As in Wald (1949), we could get a block diagonal matrix for $\Sigma_{\theta^\star}$ with $\Sigma_{12} = 0$. Such a choice is always available, at least in a neighborhood of θ_0. However, this is one choice among many and in the specific case of the ratio, it seems natural and more convenient to choose the reparametrization (3.1).

Now recall that, in a Wald statistic, a consistent estimator of the metric (or asymptotic variance) is only needed under H_0. We now exploit the information of the null hypothesis to replace $\hat{\psi}_n$ by ψ_0 in the estimation of the metric:

$$\Sigma_\psi(\hat{\psi}_n, \hat{\beta}_{1n}) \text{ is replaced by } \Sigma_\psi(\psi_0, \hat{\beta}_{1n})$$

Definition 3.1. *(Modified-Wald)*
The Modified-Wald statistic is defined as follows:

$$MW_n(\psi_0) = n \left[\hat{\psi}_n - \psi_0\right]' \Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n}) \left[\hat{\psi}_n - \psi_0\right]$$

where

$$\Sigma_\psi(\psi_0, \beta_{1n}) = \beta_{1n}^{-1} \left([-\psi_0' \; 1] \otimes I_r\right) \Sigma_\theta \left([-\psi_0' \; 1] \otimes I_r\right)' (\beta_{1n}^{-1})' \tag{3.3}$$

See the appendix for detailed calculations of Σ_ψ.

The Modified-Wald statistic $MW_n(\upsilon_0)$ can actually be reexpressed as

$$MW_n(\upsilon_0) = n\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\upsilon_0\right)' \left[\left([-\upsilon_0' \; 1] \otimes I_r\right)\Sigma_\theta\left([-\upsilon_0' \; 1] \otimes I_r\right)'\right]^{-1}\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\upsilon_0\right) \quad (3.4)$$

To conclude, our convenient reparametrization enables us to separate the parameter space into two subspaces: the first one contains the r directions *fixed* under the null hypothesis and the second one collects the remaining *free* directions. This permits to naturally incorporate the informational content of the null hypothesis in the estimation of the weighting matrix and hence reduces the dimension of the nuisance parameters when calculating the metric. Moreover, the Modified-Wald statistic (3.4) does not depend directly on the ratio. This will lead to a valid test whatever the chosen asymptotic scenario. See section 4 for some illustrations. Finally, the Modified-Wald procedure keeps the computational appeal of a classic Wald procedure in the sense that constrained estimation is avoided.

3.2 Properties

This section collects the main theoretical results of the essay. In particular, we show the asymptotic equivalence between the Modified-Wald statistic and the Wald statistic. First, we recall some definitions:

Definition 3.2. *(Power function and Consistency)*
Consider the following test, $H_0 : \theta \in \Theta(\upsilon_0)$ vs $H_1 : \theta \in \Theta/\Theta(\upsilon_0)$ with $\Theta(\upsilon_0) = \{\theta \in \Theta \; / \; \upsilon(\theta) = \upsilon_0\}$.
i) The power function of the test is the probability that the test correctly rejects the null hypothesis H_0 when the alternative H_1 is true.
ii) ((14.2) in van der Vaart (1998)) A sequence of tests with power function $\pi_n(\theta)$ is asymptotically consistent at level α against the alternative υ if the two properties below are satisfied:

$$(1) \quad \pi_n(\theta) \xrightarrow{n\to\infty} 1 \quad \forall\, \theta \in \Theta/\Theta(\upsilon_0)$$
$$(2) \quad \limsup_{n\to\infty} \sup_{\theta\in\Theta(\upsilon_0)} \pi_n(\theta) \leq \alpha$$

Theorem 3.1. *(Asymptotic Equivalence between Modified-Wald and Wald)*
i) Under H_0, the Modified-Wald statistic and the Wald statistic are asymptotically equivalent.

ii) Under a sequence of local alternatives $H_{1T} : \theta \in \Theta_n(\psi_0)$ *with* $\Theta_n(\psi_0) = \{\theta \in \Theta \ / \ \psi(\theta) = \theta_0 + \gamma/\delta_n\}$ *where* γ *is a fixed deterministic* r*-vector and* δ_n *a deterministic sequence such that* $\delta_n \xrightarrow{n} \infty$*, the Modified-Wald statistic and the Wald statistic are asymptotically equivalent.*

iii) Both associated tests are consistent.

From the above results, we can naturally define the Modified-Wald confidence region.

Corollary 3.2. *(Confidence Region)*
Under $H_0 : \psi(\theta) = \psi_0$, $MW_n(\psi_0)$ *is distributed as a Chi-square with* r *degrees of freedom and the associated confidence region with level* $(1-\alpha)$ *is defined as:*

$$CR_\psi(\alpha) = \left\{\psi_0 \ / \ MW_n(\psi_0) \leq \chi^2_\alpha(r)\right\}$$

where $\chi^2_\alpha(r)$ *denotes the* $(1-\alpha)$*-quantile of the Chi-square distribution with* r *degrees of freedom.*

We also show the validity of the Modified-Wald procedure.

Theorem 3.3. *(Validity of Modified-Wald)*

$$\inf_{\theta_0 \in \Theta} P_{\theta_0}\left[\psi(\theta) \in CR_\psi\right] > 0$$

The above result demonstrates the validity of the Modified-Wald procedure because, for any true value of the parameter $\theta_0 \in \Theta$, the coverage probability is strictly positive. This is especially true in a specific setting where the denominator gets arbitrary close to singularity.

The confidence region $CR_\psi(\alpha)$ defined in Corollary 3.2 can be expressed as follows:

$$\begin{aligned} CR_\psi(\alpha) \ = \ & \Big\{\psi_0 \in \mathbb{R}^r \ / \ n\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right)' \left[\left([-\psi_0' \ 1] \otimes I_r\right) \Sigma_\theta \left([-\psi_0' \ 1] \otimes I_r\right)'\right]^{-1} \\ & \times \left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right) \leq \chi^2_\alpha(r)\Big\} \end{aligned} \tag{3.5}$$

The above expression (3.5) can be reinterpreted as the confidence region that would have been obtained after inverting a classic Wald statistic, if performed on a linear equivalent reformulation of the null hypothesis $H_0^* : \beta_2 - \beta_1\psi_0 = 0$. This is actually known as the

Fieller principle. This has been pointed out earlier by Gregory and Veall (1985): Wald statistics are likely to be poorly approximated by standard asymptotic distributions if the constraint function is nonlinear and *not constructed in the best way*. Our context provides a good illustration: performing Wald on H_0 gives poor results whereas performing Wald on H_0^* works just fine. Note that the Modified-Wald procedure is reliable in both cases.

The confidence region (3.5) is not a typical quadric region as obtained from a classic Wald procedure. In general, it will not be possible to characterize the potential shapes of these regions. The unidimensional case ($r = 1$) is treated analytically below and the bidimensional ($r = 2$) is studied with Monte Carlo simulations in section 4.
In the unidimensional case, the above confidence region (3.5) can be rewritten as,

$$CR_\psi(\alpha) = \left\{ \psi_0 \;/\; \psi_0^2 \left[\hat{\beta}_{1n}^2 - \frac{\chi_\alpha^2(1)}{n} \sigma_{11} \right] - 2\psi_0 B + C \leq 0 \right\} \tag{3.6}$$

where $B = \hat{\psi}_n \hat{\beta}_{1n}^2 - \sigma_{12}\chi_\alpha^2(1)/n$, $C = \hat{\psi}_n^2 \hat{\beta}_{1n}^2 - \sigma_{22}\chi_\alpha^2(1)/n$ and $\Sigma = [\sigma_{ij}]_{i,j}$ is the (known) asymptotic variance of $[\hat{\beta}_{1n}\ \hat{\beta}_{2n}]'$. In the unidimensional case, the confidence region is simply a quadratic region. Hence, we can exhaustively describe its potential shapes:

i) an interval;
ii) the entire real line;
iii) an empty set;
iv) a disjoint union of two semi-intervals.

We can also calculate the probability of having an unbounded region. Recall that a quadratic region like (3.6) is unbounded whenever the coefficient of ψ_0^2 is negative.

Proposition 3.4. *(Probability of an Unbounded Region)*

$$P\left(CR_\psi(\alpha) \text{ unbounded}\right) = P\left[n\frac{\hat{\beta}_1^2}{\sigma_{11}} < \chi_\alpha^2(1) \right] \xrightarrow{n\to\infty} 1 - \alpha$$

when the true (unknown) value of β_1 is arbitrarily close to 0, that is when we consider sequences of true parameters converging to 0.

The region (3.5) is unbounded whenever we cannot reject the hypothesis $\beta_1 = 0$. The above probability converges to the upper bound of the necessary condition of Dufour (1997). Finally, note that the region (3.5) corresponds to the confidence region derived by Dufour

3.3 An alternative interpretation

In this subsection, we attempt to reinterpret the rather imprecise statement *when β is arbitrary close to zero* from equation (3.4). So far, in the spirit of Dufour (1997), we have considered identification issues happening at the frontier of the parameter space. And we were able to define a valid inference procedure, the Modified Wald test, associated with (potentially unbounded) confidence regions. Now, we connect the above (concrete) identification issue to another (well-known) scenario where the parameter value is artificially linked to the sample size (see for instance Staiger and Stock (1997)). Following Antoine and Renault (2007), our setup is reinterpreted as follows:

Assumption 2. *(Reinterpretation of Assumption 1)*

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{1n} - \frac{1}{n^{\lambda(\theta)}}\beta_{1,0} \\ \hat{\beta}_{2n} - \beta_{2,0} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta) \text{ with } 0 \leq \lambda(\theta) \leq 1/2 \text{ and for some fixed matrix } \beta_{1,0}$$

This setup permits to recover a couple of interpretations of the literature on weak instruments: i) when $\inf_{\theta\in\Theta} \lambda(\theta) = 1/2$, we have *weak identification*, as in Stock and Wright (2000); ii) when $0 < \inf_{\theta\in\Theta} \lambda(\theta) < 1/2$, we have *nearly-weak identification*, as in Hahn and Kuersteiner (2002); iii) and when $\inf_{\theta\in\Theta} \lambda(\theta) = 0$, we have standard identification.

The confidence region is the same as (3.6): only the interpretation of the behavior of the parameter is altered. Let us now calculate the probability of getting an unbounded confidence region:

$$P\left[n\frac{\hat{\beta}_{1n}^2}{\sigma_{11}} < \chi_\alpha^2(1)\right] = P\left[\sqrt{n}\left(\hat{\beta}_{1n} - \frac{1}{n^{\lambda(\theta)}}\beta_{1,0}\right) + n^{1/2-\lambda(\theta)}\beta_{1,0} < \sigma_{11}\sqrt{\chi_\alpha^2(1)}\right]$$

Since

$$\sqrt{n}\left(\hat{\beta}_{1n} - \frac{1}{n^{\lambda(\theta)}}\beta_{1,0}\right) = \mathcal{O}_P(1),$$

we get:

- Strong and Nearly-weak cases: $0 \leq \inf_{\theta\in\Theta} < 1/2$ $\quad P(CR_\psi(\alpha) \text{ unbounded}) \xrightarrow{n} 0$
- Weak case: $\inf_{\theta\in\Theta} = 1/2$ $\quad P(CR_\psi(\alpha) \text{ unbounded}) \xrightarrow{n} 1-\alpha$

To conclude, in terms of unbounded confidence region, the (artificial) weak identification case replicates the behavior of the more concrete situation where we get arbitrary close to

the frontier of the non-identification subset. To our knowledge, this is the first time such a parallel is made.

4 (Nearly)-Weak Identification Applications

In this section, we apply the previous methods and results to two econometric settings. First, a simulation exercise is designed to compare the Wald and Modified-Wald procedures in the bidimensional setting of section 3.2. The second application focuses on the well-known linear single-equation instrumental variables regression model.

4.1 Ratio of parameters

In the multidimensional case ($r > 1$), the possible shapes of the confidence region (3.5) are hard to list since it does not belong to any known class of regions. We then perform a simulation exercise in the bidimensional case ($r = 2$) to compare the Wald and Modified-Wald procedures through *averaged confidence regions* (to be defined) and power functions.

We consider the following random vector, as in assumption 1:

$$\sqrt{n}\left(vec[\hat{\beta}_{1n}\ \hat{\beta}_{2n}] - vec[\beta_{1,0}\ \beta_{2,0}]\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta)$$

for some known positive-definite nonsingular matrix Σ_β and $vec[\beta_{1,0}\ \beta_{2,0}] = [b\ 0\ 0\ 1\ 1\ 1]'$, where b is a real number that depends on how close to singularity the matrix $(\beta_{1,0})$ is. Typically, we consider three cases of interest: $b = 1;\ 0.1;\ 0.01$.

Practically, we simulate a random sample of size n for the (6,1)-random vector $vec(\beta)$ and we use the sample mean as a root n consistent estimator. We then use a bidimensional grid to build a confidence region for the (2,1) ratio of parameter, $\psi = \beta_1^{-1}\beta_2$. Comparing this region to the one obtained by inverting the Wald-type statistic would be the result of one random sample only. Hence, even though it would provide an easy way to compare visually the two procedures, it might be the case that another sample leads to a different region (different shape and/or different properties). This is the reason why we decide to produce averaged confidence regions, or ACR. These regions are based on M (usual) confidence regions, each of them built for M different samples: the ACR collects the points that appear in at least $q\%$

of the M confidence regions, for some chosen q. ACR may have irregular shape but they reflect more accurately what the procedure does, hence permitting to compare reliably both procedures. We consider here a sample of size 200 and 1000 replications. In addition, we also provide some results on power functions, which are as well obtained through many different samples.

First, we consider a benchmark case without identification issues where $b = 1$. The Wald and Modified-Wald procedures perform pretty similarly. See figure (III.1) for the 85%-ACR and figures (III.6) and (III.7) for the power curves, respectively the Modified-Wald and Wald procedures. The averaged confidence regions are bounded ellipsoid, relatively narrow and roughly centered around the true parameter $\theta_0 = (1, 1)$. The power curves show a well around the true parameter, where the power to reject the null falls below 10%. Everywhere else, both methods have a power very close to 100%.

Second, we consider a case with mild identification issues where $b = .1$. The Wald and Modified-Wald procedures do not yield anymore to similar conclusions. As for the ACR, the Modified-Wald procedure still proposes a bounded ellipsoid but a lot larger than in the previous case. See the 85%-ACR figure (III.2). Notice that Wald 85%-ACR is empty. In fact, we have to go as low as $q = 16\%$ to get a non-empty region. See figure (III.3) for the 15%-ACR. The Wald procedure is pretty unstable since its output varies a lot from one sample to another. Moreover the scales of the regions are very different: the Wald procedure still proposes some relatively narrow region while the Modified-Wald region has become pretty large: this can be interpreted as a lack of information. As for the power curves, the Wald procedure clearly over-rejects: its power does not go below 85%. This is of course directly linked to the above averaged confidence region. In comparison, the Modified-Wald procedure performs fairly well: it still presents the well-shape around the true value of the parameter, with a power decreasing below 20%. Note that there is power in every direction, with a power increasing above 90%. It is slightly asymmetric, with less power towards large values of the first component of θ and low values of the second component of θ. See figures (III.8) and (III.9).

Finally, we consider a case with more serious identification issues where $b = .01$. Many of the facts pointed out in the mild case of weak identification are now exaggerated. The Modified-Wald procedure proposes an unbounded confidence region: the outside of an hyperboloid which still contains the true value of the parameter. The unboundedness of the confidence

region allows us to conclude that the sample information is not sufficient to provide sharp inference on the parameter of interest. See figure (III.4) for the 75%-ACR. As expected after the above case with mild identification issues, the Wald 75%-ACR is empty. We have to go as low as $q = 2\%$ to get a non-empty Wald ACR. See figure (III.5) for the 1%-ACR. As for the power curves, again the Wald procedure cannot discriminate much: the power does not go below 97%. The Modified-Wald procedure proposes a butterfly-shape power, which is natural after the hyperbolic unbounded confidence region: the power is close to 100% outside the hyperbola (butterfly-shape) and it falls below 15% inside. Again, the shape of the power function should be understood as a lack of sample information to discriminate among different subsets of parameter values. See figures (III.10) and (III.11).

4.2 Application to the Single-equation IV model

Consider the following just-identified structural model:

$$\begin{cases} \underset{(n,1)}{y} = \underset{(n,r)}{Y}\underset{(r,1)}{\psi} + \underset{(n,1)}{u} \\ \underset{(n,r)}{Y} = \underset{(n,r)}{X}\underset{(r,r)}{\Pi} + \underset{(n,r)}{V} \end{cases} \tag{4.1}$$

where Y is an endogenous variable, X a strictly exogenous (instrumental) variable, ψ an unknown coefficients vector and Π an unknown coefficients matrix and $[u\ V]'$ a matrix of homoscedastic errors.

We are interested in providing inference on the structural parameter ψ, which is identifiable if and only if the rank of the matrix Π is full ($Rank\ \Pi = r$). A case of interest appears when Π is (potentially) close to singularity. In the spirit of Staiger and Stock (1997), and more recently Hahn and Kuersteiner (2002), the matrix Π is (artificially) linked to the sample size n as follows:

$$\Pi = \frac{C}{n^{\lambda}} \text{ where } C \text{ is a fixed deterministic matrix of rank } r \text{ and } 0 \leq \lambda \leq 1/2.$$

This framework has been referred to as *valid instruments* when $\lambda = 0$ (Π is a full-rank deterministic matrix), to *nearly-weak instruments* when $0 < \lambda < 1/2$, to *weak instruments* when $\lambda = 1/2$ and to *invalid instruments* when $\Pi = 0$. Hence, this setting considers a

sequence of models in which the moment condition $E(u_i Z_i)$ depends on the sample size n. However, this is just a convenient (artificial) device. This is related to the concept of *drifting DGP* used to study the local asymptotic power of tests (see for instance Davidson and MacKinnon (1993)).

The classical hypothesis associated to the study of the model (4.1) are presented next.

Assumption 3.

$$i) \quad \frac{1}{n}[u\ V]'[u\ V] \xrightarrow{P} \Sigma \text{ with } \Sigma = \begin{pmatrix} \sigma_u & \sigma_{uV} \\ \sigma_{Vu} & \sigma_V \end{pmatrix} \text{ symmetric and positive definite}$$

$$ii) \quad \frac{1}{n}X'X \xrightarrow{P} Q \text{ symmetric and positive definite}$$

$$iii) \quad \frac{1}{\sqrt{n}}X'[u\ V] \xrightarrow{d} \Psi = [\psi_{Xu}\ \psi_{XV}] \text{ where } vec[\psi_{Xu}\ \psi_{XV}] \sim \mathcal{N}(0, \Sigma \otimes Q)$$

The well-known two-step least squares (2-SLS) estimator of ψ (in our just-identified case) is defined as: $\tilde{\psi}_n = (X'Y)^{-1}X'y$. We now recall a useful result from Hahn and Kuersteiner (2002). It justifies that the general setting of section 3 applies and, in particular, that Assumption 1 is satisfied.

Proposition 4.1. *(Hahn and Kuersteiner (2002))*

i) For $0 \leq \lambda < 1/2$,

$$n^{1/2-\lambda}\left[\tilde{\psi}_n - \psi_0\right] \xrightarrow{P} \mathcal{N}(0, \Sigma)$$

$$\text{where } \Sigma = \sigma_u \times \left[plim\left(\frac{Y'X}{n}\right)\left(plim\frac{X'X}{n}\right)^{-1}\left(plim\frac{X'Y}{n}\right)\right]^{-1}$$

ii) For $\lambda = 1/2$,

$$vec\left[\frac{X'Y}{\sqrt{n}}\ \frac{X'y}{\sqrt{n}}\right] \xrightarrow{d} \mathcal{N}\left(vec[QC\ QC\psi_0], \Sigma_y(\psi_0)\right)$$

$$\text{where } \Sigma_y = \begin{pmatrix} \sigma_V & \sigma_{Vu} + \sigma_V\psi_0 \\ \sigma_{uV} + \psi_0'\sigma_V & \sigma_u + \psi_0'\sigma_V\sigma_0 + 2\sigma_{uV}\psi_0 \end{pmatrix} \otimes Q$$

iii) With invalid instruments, the above result remains true after replacing C by the null matrix.

Definition 4.3. *(Modified-Wald)*
In the context of model (4.1), the Modified-Wald statistic writes:

$$MW_n(\psi_0) = n\left[\hat{\psi}_n - \psi_0\right]' \Sigma_\psi(\psi_0)^{-1}\left[\hat{\psi}_n - \psi_0\right]$$

with $\Sigma_\psi(\psi_0) = (y - Y\psi_0)'(y - Y\psi_0) \times [Y'X(X'X)^{-1}X'Y]^{-1}$

We can now state the main result of this section:

Theorem 4.2. *(Modified-Wald)*
Under H_0 and some regularity conditions, $MW_n(\psi_0)$ is asymptotically distributed as a Chi-square with r degrees of freedom.

It is important to note that the statistic $MW_n(\psi_0)$ can be rewritten as follows,

$$MW_n(\psi_0) = n\frac{(y - Y\psi_0)'X(X'X)^{-1}X'(y - Y\psi_0)}{(y - Y\psi_0)'(y - Y\psi_0)}$$

The above expression does not depend on $(X'Y)^{-1}$. This implies that the Modified-Wald procedure is well defined for any asymptotic scenario, strong, nearly-weak or weak instruments. This is another sharp contrast with the Wald procedure.

The confidence region is then defined as:

$$CR_\psi(\alpha) = \left\{\psi_0 \,/\, n\left[\hat{\psi}_n - \psi_0\right]' \Sigma_\psi^{-1}(\psi_0, \hat{\beta}_1)\left[\hat{\psi}_n - \psi_0\right] \leq \chi^2_\alpha(r)\right\}$$

where $\chi^2_\alpha(r)$ denotes the $(1-\alpha)$-quantile of a Chi-square distribution with r degrees of freedom. The following useful reformulation of the confidence region,

$$CR_\psi(\alpha) = \{\psi_0 \,/\, \psi_0' Y'AY\psi_0 - 2\psi_0' Ay + y'Ay \leq 0\} \quad \text{with} \quad A = P_X - \chi^2_\alpha(r) I_n/n$$

emphasizes it is a (classic) quadric confidence region. As the multidimensional extension of the quadratic case, the region $CR_\psi(\alpha)$ is unbounded if and only if the matrix $Y'AY$ is negative definite. Hence, we have[5]:

$$P(CR_\psi(\alpha)\ unbounded) = P\left[(P_X - \chi^2_\alpha(r)/n\, I_n) << 0\right]$$

[5]Recall that for for any 2 (r,r)-matrices M_1 and M_2, we have:

$$M_1 >> M_2 \ iff\ u'M_1u \leq u'M_2u \ \text{ for any r-vector } u$$

Other methods have yielded to a similar structure of the confidence region but with a different matrix A. Here are some examples:

Method	Matrix A
Modified Wald	$P_X - \chi^2_\alpha(r) I_n/n$
Anderson-Rubin	$P_X - M_X \chi^2_\alpha(r)/(n-r)$
Wang and Zivot (GMM0)	$P_X Y'(Y' P_X Y')^{-1} Y' P_X - I_n \chi^2_\alpha(r)/n$
Wang and Zivot (LR-LIML)	$I_n - k(LIML) \exp[\chi^2_\alpha(r)/n] M_X$

It is difficult to precisely compare their associated confidence regions. However, the Modified-Wald confidence region allows for some natural and intuitive additional results:

Proposition 4.3. *(Unbounded Confidence Region in the Unidimensional case)*
In the unidimensional case ($r = 1$), we have:

i) *For* $0 \leq \lambda < 1/2$: $P(CR_\psi(\alpha)$ *unbounded*$) \xrightarrow{n\to\infty} 0$

ii) *For* $\lambda = 1/2$: $P(CR_\psi(\alpha)$ *unbounded*$) \xrightarrow{n\to\infty} 1 - \tilde{\alpha}$ *with* $\tilde{\alpha} > \alpha$

iii) *With unidentification:* $P(CR_\psi(\alpha)$ *unbounded*$) \xrightarrow{n\to\infty} 1 - \alpha$

With strong identification, the confidence region is almost surely bounded. In the opposite scenario, with unidentification, the probability of an unbounded confidence region converges towards the upper bound derived by Dufour (1997) in his necessary condition (see also section 2). Now, in the intermediate case of weak identification, this probability tends to some real number between 0 and the upper bound, which depends on the asymptotic value of $X'Y/\sqrt{n}$. See also appendix for further details. It is also interesting to point out that, in this case, nearly-weak identification behaves similarly to strong identification: in some sense, nearly-weak identification is *not weak enough* to lead to unbounded confidence regions. This result can actually be extended to the multidimensional case:

Proposition 4.4. *(Unbounded Confidence Region in the Multidimensional case)*
For $0 \leq \lambda < 1/2$: $P(CR_\psi(\alpha)$ *unbounded*$) \xrightarrow{n\to\infty} 0$

The difference between the strong and the nearly-weak identification cases is the rate of the above convergence: with strong instruments, the rate corresponds to the fastest available one, $(1/n)$; with nearly-weak instruments, this rate is only $(1/n^{1-2\lambda})$ and decreases with weaker instruments.

Finally, we might also be interested in providing inference on a subvector of the structural parameter θ. Without loss of generality, we suppose that inference is provided about the first m components of θ denoted as ${}_m\theta$. The classical maintained assumption in the literature is to suppose that the remaining components (the ones not involved in H_0 and denoted as ${}_{r-m}\theta$) are strongly identified. In such a context, we can show that all the previous results remain valid when ${}_{r-m}\theta$ are simply replaced by their respective consistent estimators.

5 Conclusion

In this chapter, we proposed a new inference method, the Modified-Wald procedure, to provide reliable inference about a multidimensional ratio of parameters. This new method is based on the Wald statistic and shows the same computational tractability. In addition, it provides unbounded confidence regions when the identification fails, as suggested by Dufour (1997). The key idea consists in integrating the informational content of the null hypothesis of interest to compute its metric.

We have shown that the Modified-Wald statistic is asymptotically equivalent to the Wald statistic. The associated confidence region with level α is unbounded with a probability as high as (1-α), the upper bound suggested in Dufour (1997). These results were applied to two examples. In the first one, a simulation exercise is designed to compare the properties of the Wald and Modified-Wald procedures with a bidimensional ratio. Generally speaking, the Modified-Wald behaves pretty well. When there is no identification issue, the confidence region is as narrow as the Wald one (recall that Wald is perfectly valid in such a case). And, when there is a serious case of identification issues, the Modified-Wald confidence region is unbounded, contrary to the (invalid) Wald one which remains narrow and bounded. Our method is then able to detect insufficient sample information to provide sharp inference.
The second application focuses on the well-known linear single-equation instrumental variables regression model. In this context, the identification issues are modeled after artificially linking the parameter value to the sample size. We can distinguish between three cases of interest depending on how fast identification is lost: i) with strong and nearly-weak identification, the confidence region is almost surely bounded; ii) with unidentification, it is asymptotically unbounded with a probability equal to the upper bound; iii) with weak identification, it is asymptotically unbounded with a smaller probability.

These results are promising because they build up a connection between the identification loss at the frontier of the parameter space and its (artificial) econometric modelization through the sample size device.

Appendix

Proofs of the main results

Proof of Equation (3.3) *(Asymptotic Variance of the Ratio)*:
From equation (2.3),

$$\Sigma_\psi(\hat{\theta}_n) = \frac{\partial\psi(\hat{\theta}_n)}{\partial\theta'}\Sigma_\theta\frac{\partial\psi'(\hat{\theta}_n)}{\partial\theta}$$

So we only need to calculate:

$$\frac{\partial\psi(\theta)}{\partial\theta'} = \frac{\partial\psi(\theta)}{\partial vec[\beta_1\ \beta_2]'} = \left[\frac{\partial\psi(\theta)}{\partial vec(\beta_1)'}\ \frac{\partial\psi(\theta)}{\partial\beta_2'}\right]$$

It is well known that:

$$\frac{\partial\psi(\theta)}{\partial\beta_2'} = \beta_1^{-1}$$

To derive the second part of the calculation, we first recall some useful formulae[6]:

$$\begin{aligned}
&i) \quad \frac{\partial vec(\beta_1^{-1})}{\partial vec(\beta_1)'} = -(\beta_1^{-1})' \otimes \beta_1^{-1} \\
&ii) \quad \frac{\partial\beta_1^{-1}}{\partial t} = -\beta_1^{-1}\frac{\partial\beta_1}{\partial t}\beta_1^{-1} \quad \text{where } t \text{ is a scalar} \\
&iii) \quad \beta_1^{-1}\beta_2 = I_r\beta_1^{-1}\beta_2 = vec(I_r\beta_1^{-1}\beta_2) = (\beta_2' \otimes I_r)vec(\beta_1^{-1})
\end{aligned}$$

We can then derive:

$$\begin{aligned}
\frac{\partial(\beta_1^{-1}\beta_2)}{\partial vec(\beta_1)'} = \frac{\partial vec(\beta_1^{-1}\beta_2)}{\partial vec(\beta_1)'} &= (\beta_2' \otimes I_r)\frac{\partial vec(\beta_1^{-1})}{\partial vec(\beta_1)'} \\
&= -(\beta_2' \otimes I_r)((\beta_1^{-1})' \otimes \beta_1^{-1}) \\
&= -(\beta_2'(\beta_1^{-1})' \otimes \beta_1^{-1}) \\
&= -(\psi' \otimes \beta_1^{-1})
\end{aligned}$$

Hence, $\dfrac{\partial\psi}{\partial vec[\beta_1\ \beta_2]'} = [-\psi' \otimes \beta_1^{-1}\ ,\ \beta_1^{-1}] = [-\psi' \otimes \beta_1^{-1}\ ,\ 1 \otimes \beta_1^{-1}] = [-\psi'\ ,\ 1] \otimes \beta_1^{-1}$

The commas emphasize the partitioned structure of the matrix. This leads to the expected result. ■

[6] See for instance Abadir and Magnus (2005) chapter 13.

Proof of Theorem 3.1 *(Asymptotic Equivalence between Modified-Wald and Wald)*:
i) and ii) Recall the following,

$$\begin{aligned} W_n &= n\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right)' A^{-1}(\hat{\psi}_n)\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right) \\ MW_n &= n\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right)' A^{-1}(\psi_0)\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right) \end{aligned}$$

with $A(\psi) = ([-\psi' \; 1] \otimes I_r)\,\Sigma_\theta\,([-\psi' \; 1] \otimes I_r)'$. Since $\sqrt{n}\left(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi_0\right) = \mathcal{O}_P(1)$, it is enough to show that:

$$\begin{aligned} A^{-1}(\hat{\psi}_n) - A^{-1}(\psi_0) \xrightarrow{P} 0 \quad &\Leftrightarrow \quad A^{-1}(\hat{\psi}_n)\left[A(\psi_0) - A(\hat{\psi}_n)\right] A^{-1}(\psi_0) \xrightarrow{P} 0 \\ &\Leftrightarrow \quad A(\hat{\psi}_n) - A(\psi_0) \xrightarrow{P} 0 \\ &\Leftrightarrow \quad \left([(\psi_0 - \hat{\psi}_n)\; 0] \otimes I_r\right)\Sigma_\theta\left([(\psi_0 - \hat{\psi}_n)\; 0] \otimes I_r\right)' \xrightarrow{P} 0 \\ &\Leftrightarrow \quad [(\psi_0 - \hat{\psi}_n)'\; 0] \xrightarrow{P} 0 \;\text{ since } \Sigma_\theta \text{ positive definite} \end{aligned}$$

The last convergence result is true both under H_0 and a sequence of local alternatives.

iii) The test $H_0 : \psi(\theta) = \psi_0$ vs $\psi(\theta) \neq \psi_0$ can be rewritten as, $H_0(\psi_0) : \theta \in \Theta(\psi_0)$ vs $\theta \in \Theta/\Theta(\psi_0)$ where $\Theta(\psi_0) = \{\theta \in \Theta \;/\; \psi(\theta) = \psi_0\}$. We need to verify the two statements of definition 3.2ii).
- Proof of (1) (from definition 3.2 ii)):

$$\forall\, \theta \in \Theta - \Theta(\psi_0) \;\; \pi_n(\theta) = P_\theta\left(n(\hat{\psi}_n - \psi_0)'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})(\hat{\psi}_n - \psi_0) \geq \chi^2_\alpha(r)\right)$$

Note that:

$$\begin{aligned} n(\hat{\psi}_n - \psi_0)'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})(\hat{\psi}_n - \psi_0) = \; & n(\hat{\psi}_n - \psi(\theta))'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})(\hat{\psi}_n - \psi(\theta)) \;\Big\}(a) \\ & + 2n(\psi(\theta) - \psi_0)'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})(\hat{\psi}_n - \psi(\theta)) \\ & + n(\psi(\theta) - \psi_0)'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})(\psi(\theta) - \psi_0) \;\Big\}(b) \end{aligned}$$

$$(a) \;=\; \sqrt{n}(\hat{\psi}_n - \psi(\theta))'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})\sqrt{n}(\hat{\psi}_n - \psi(\theta)) \xrightarrow{d} Z'S^{-1}Z$$

where under the data generating process with parameter θ (denoted as DGP(θ)), $Z \sim \mathcal{N}(0, \Sigma_\psi)$ and $\Sigma_\psi(\psi_0, \hat{\beta}_{1n}) \xrightarrow{P} \Sigma_\psi(\psi_0, \beta_{1,0}) = S$ a symmetric positive definite matrix.

$$\begin{aligned} (b) \;&=\; n(\psi(\theta) - \psi_0)'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})(\hat{\psi}_n - \psi(\theta) + \hat{\psi}_n - \psi_0) \\ &=\; \sqrt{n}(\psi(\theta) - \psi_0)'\Sigma_\psi^{-1}(\psi_0, \hat{\beta}_{1n})\sqrt{n}(\hat{\psi}_n - \psi(\theta) + \hat{\psi}_n - \psi_0) \end{aligned}$$

and under DGP(θ)

$$\Sigma_{\psi}^{-1}(\psi_0, \hat{\beta}_{1n})\sqrt{n}\left(\hat{\psi}_n - \psi(\theta) + \hat{\psi}_n - \psi_0\right) \overset{a}{\sim} S^{-1}\left[Z + \sqrt{n}\left(\psi(\theta) - \psi_0\right)\right]$$

$$\text{and} \quad \sqrt{n}\left(\psi(\theta) - \psi_0\right)' \overset{n\to\infty}{\longrightarrow} +\infty$$

Hence we deduce that, $\pi_n(\theta) \overset{n\to\infty}{\longrightarrow} 1$

- Proof of (2) (from definition 3.2 ii)):

$$\forall\theta \in \Theta(\psi_0) \text{ and fixed } n, \quad \pi_n(\theta) \quad = \quad P_\theta\left(MW_n(\psi(\theta)) > \chi^2_\alpha(r)\right)$$

Since under H_0, $MW_n(\psi(\theta)) \xrightarrow{d} C(r)$ (where $C(r)$ denotes the Chi-square distribution with r degrees of freedom) and using the continuity of $x \to P_\theta(MW_n(\psi(\theta)) \leq x)$ at all x, we deduce that $\pi_n(\theta) \overset{n\to\infty}{\longrightarrow} \alpha$. ■

Proof of Theorem 3.3 *(Validity of Modified-Wald)*:
Recall the following,

$$\begin{aligned} & P_{\theta_0}\left(\psi(\theta) \in CR_\psi(\alpha)\right) \\ = \; & P_{\theta_0}\left(n[\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi]'\left[([-\psi' \; 1] \otimes I_r)\Sigma_\theta([-\psi' \; 1] \otimes I_r)'\right]^{-1}[\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi] \leq \chi^2_\alpha(r)\right) \end{aligned}$$

with $\sqrt{n}[vec(\hat{\beta}_{1n} - \beta_{1,0})' \; (\hat{\beta}_{2n} - \beta_{2,0})']' \xrightarrow{d} \mathcal{N}(0, \Sigma_\theta)$.
Define the following matrix $V = \left[([-\psi' \; 1] \otimes I_r)\Sigma_\theta([-\psi' \; 1] \otimes I_r)'\right]^{-1}$. We deduce that:

$$n(\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi)'V[\hat{\beta}_{2n} - \hat{\beta}_{1n}\psi] \sim NC\left(r, \lambda\right)$$

where $NC(r, \lambda)$ denotes the noncentral Chi-square distribution with r degrees of freedom and noncentrality parameter λ. Here, $\lambda = \|\beta_{2,0} - \beta_{1,0}\psi\|_V$.

Then, $\inf_{\theta_0 \in \Theta} P_{\theta_0}\left(\psi(\theta) \in CR_\psi(\alpha)\right) = \inf_{\theta_0 \in \Theta} P_{\theta_0}\left(NC(r, \|\beta_{2,0} - \beta_{1,0}\psi\|^2) \leq \chi^2_\alpha(r)\right) > 0$

because $\|\beta_{2,0} - \beta_{1,0}\psi\|^2_V = \|\beta_{1,0}\|^2_V \times \|\psi_0 - \psi\|^2_V$ is bounded. ■

Proof of Proposition 3.4 *(Probability of an Unbounded Region)*:
Directly follows from the asymptotic normality of $\hat{\beta}_1$. ■

Proof of Theorem 4.2 *(Modified-Wald)*:
i) Consider the case $0 \leq \lambda < 1/2$, using proposition 4.1, we only need to prove that $\Sigma_\psi(\psi_0)$

Now we have:

$$\begin{aligned}\Sigma_v(\psi_0) &= n^{1-2\lambda}\frac{(y-Y\psi_0)'(y-Y\psi_0)}{n}[Y'X(X'X)^{-1}X'Y]^{-1} \\ &= \frac{(y-Y\psi_0)'(y-Y\psi_0)}{n}\left[\left(\frac{Y'X}{n^{1-\lambda}}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'Y}{n^{1-\lambda}}\right)\right]^{-1}\end{aligned}$$

We now recall a useful lemma from Hahn and Kuersteiner (2002):

Lemma A. *(Hahn and Kuersteiner (2002))*
For $0 \leq \lambda < 1/2$,

$$\frac{1}{n^{1-\lambda}}X'Y \sim QC + \frac{1}{n^{1/2-\lambda}}v_{XV} = QC + \mathcal{O}_P\left(\frac{1}{n^{1/2-\lambda}}\right)$$

For $\lambda = 1/2$,

$$\frac{1}{n^{1/2}}X'Y \sim QC + v_{XV} = QC + \mathcal{O}_P(1)$$

From this lemma, $X'Y/n^{1-\lambda} \sim QC + \mathcal{O}_P(1/n^{1/2-\lambda})$. By hypothesis, $X'X/n \xrightarrow{P} Q$. Hence:

$$\left(\frac{Y'X}{n^{1-\lambda}}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'Y}{n^{1-\lambda}}\right) \xrightarrow{P} C'QC$$

σ_u can be consistently estimated under H_0 by $u(\psi_0)'u(\psi_0)/n$ where $u(\psi) = y - Y\psi$. So $\Sigma_v(\psi_0)$ is a consistent estimator of Σ_v under H_0.

ii) Consider the case $\lambda = 1/2$: we can directly applied the general theory of Modified-Wald statistic, by noting that the variance is:

$$\Sigma_v(\psi_0) = \Sigma_v(\psi_0, \hat{\beta}_{1n}) = \hat{\sigma}_u(\psi_0) \times \hat{\beta}_{1n}^{-1\prime}\hat{Q}\hat{\beta}_{1n}^{-1}$$

where $\hat{\beta}_{1n} = X'Y/n$ with $\sqrt{n}vec(\hat{\beta}_{1n} - \beta_{1,0}) \xrightarrow{d} \mathcal{N}(0, \sigma_V \otimes Q)$.
Also,

$$\hat{\sigma}_u(\psi_0) = \frac{(y-Y\psi_0)'(y-Y\psi_0)}{n} \quad \text{and} \quad \hat{Q} = \frac{X'X}{n}$$

The Modified-Wald statistic can then be rewritten as:

$$\begin{aligned} MW_n(v_0) &= n\left[\hat{v}-v_0\right]'\hat{\Sigma}^{-1}(v_0)\left[\hat{v}-v_0\right] \\ &= n\left[(X'Y)^{-1}X'y-v_0\right]'\frac{\left[(Y'X)^{-1}X'X(X'Y)^{-1}\right]^{-1}}{\hat{\sigma}_u(v_0)}\left[(X'Y)^{-1}X'y-v_0\right] \\ &= n\frac{u(v_0)'X(X'X)^{-1}X'u(v_0)}{u(v_0)'u(v_0)} \\ &\xrightarrow{d} Z_n'Z_n \sim \text{Chi-square(r)} \end{aligned}$$

with $Z_n = Q'^{-1/2}v_{Xu}$.■

Proof of Propositions 4.3 and 4.4 *(Unbounded Confidence Region)*:

$$P(CR_v(\alpha)\ unbounded) = P(Y'AY << 0) = P\left(Y'X(X'X)^{-1}X'Y << \chi^2_\alpha(r)Y'Y\right)$$

i) Consider the case $0 \leq \lambda < 1/2$.

$$P(CR_v(\alpha)\ unbounded) = P\left[\left(\frac{Y'X}{n^{1-\lambda}}\left(\frac{X'X}{n}\right)^{-1}\frac{X'Y}{n^{1-\lambda}}\right) << \chi^2_\alpha(r)\frac{Y'Y}{n^{2-2\lambda}}\right]$$

As mentioned in the previous proof, the LHS converges in probability towards some symmetric positive definite matrix $C'QC$. The RHS:

$$\begin{aligned} \frac{Y'Y}{n^{2-2\lambda}} &= \frac{1}{n^{2-2\lambda}}\left[\left(\frac{XC}{n^\lambda}+V\right)'\left(\frac{XC}{n^\lambda}+V\right)\right] \\ &= \frac{1}{n}C'\frac{X'X}{n}C + \frac{1}{n^{3/2-\lambda}}C'\frac{X'V}{n^{1/2}} + \frac{1}{n^{3/2-\lambda}}\frac{V'C}{n^{1/2}}C + \frac{1}{n^{1-2\lambda}}\frac{V'V}{n} \\ &= \mathcal{O}_P\left(\frac{1}{n}\right) + \mathcal{O}_P\left(\frac{1}{n^{3/2-\lambda}}\right) + \mathcal{O}_P\left(\frac{1}{n^{1-2\lambda}}\right) \\ &\xrightarrow{P} 0 \end{aligned}$$

because by assumption we have $X'X/n \xrightarrow{P} Q$, $X'V/n^{1/2} = v_{XV} = \mathcal{O}_P(1)$, $V'V/n \xrightarrow{P} 0$ and $0 \leq \lambda < 1/2$.

ii) Consider $\lambda = 1/2$:

$$\begin{aligned} Y'\left(P_X - \frac{\chi^2_\alpha(r)}{n}I_n\right)Y << 0 \quad &\Leftrightarrow \quad Y'P_XY << \frac{\chi^2_\alpha(r)}{n}Y'Y \\ &\Leftrightarrow \quad \frac{Y'X}{\sqrt{n}}\left(\frac{X'X}{n}\right)^{-1}\frac{X'Y}{\sqrt{n}} << \chi^2_\alpha(r)\frac{Y'Y}{n} \end{aligned}$$

Recall that we have:

$$\frac{X'Y}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(QC, \sigma_V \otimes Q)$$

and $(X'X)/n \xrightarrow{P} Q$ symmetric positive definite.
Note that we also have:

$$\begin{aligned}\frac{Y'Y}{n} &= \frac{1}{n}[\Pi'X'X\Pi + V'X\Pi + \Pi'X'V + V'V] \\ &= \frac{1}{n}\left[C'\frac{X'X}{n}C + \frac{V'X}{\sqrt{n}} + C'\frac{X'V}{\sqrt{n}} + V'V\right] \\ &\xrightarrow{P} 0 + 0 + 0 + \sigma_V\end{aligned}$$

because again $X'V/\sqrt{n} \xrightarrow{d} \mathcal{N}(0, \sigma_V \otimes Q)$ and so $X'V/n \xrightarrow{P} 0$.
In the unidimensional case, we have:

$$\begin{aligned}P\left(CR_\psi(\alpha)\ unbounded\right) &= P\left(\left(\frac{X'Y}{\sqrt{n}}\right)^2 \times \frac{1}{\frac{X'X}{n} \times \frac{Y'Y}{n}} \leq \chi^2_\alpha(1)\right) \\ &\xrightarrow{n} P\left(NC(1,\delta) \leq \chi^2_\alpha(1)\right) \\ &= 1 - \tilde{\alpha} < 1 - \alpha\end{aligned}$$

where δ represents the noncentrality parameter: its explicit formula is not needed here. The last equality comes from the comparison of a non-central Chi-square distribution with a (central) Chi-square quantile. The non-centrality parameter which shifts the distribution towards the right is due to $X'Y/\sqrt{n} \xrightarrow{d} \mathcal{N}(QC, \sigma_V Q)$.

iii) Consider the invalid instruments case: in the unidimensional case, we can produce a similar analysis,

$$\begin{aligned}P\left(CR_\psi(\alpha)\ unbounded\right) &= P\left(\left(\frac{X'Y}{\sqrt{n}}\right)^2 \times \frac{1}{\frac{X'X}{n} \times \frac{Y'Y}{n}} \leq \chi^2_\alpha(1)\right) \\ &\xrightarrow{n} P\left(C(1) \leq \chi^2_\alpha(1)\right) \\ &= 1 - \alpha\end{aligned}$$

where the last equality comes from the legitimate comparison of a (central) Chi-square distribution with a (central) Chi-square quantile. The non-centrality parameter disappears because we have $X'Y/\sqrt{n} \xrightarrow{d} \mathcal{N}(0, \sigma_V Q)$. ■

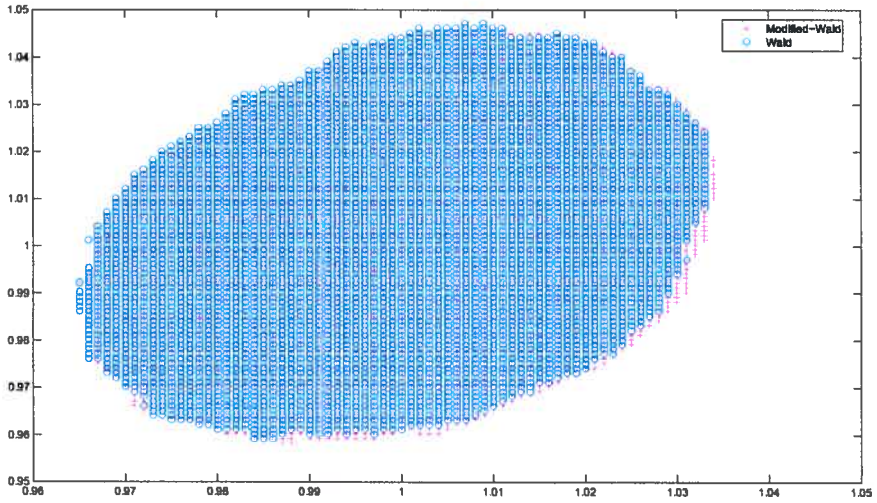

Figure III.1: 85%-Averaged Confidence Region when b=1 using Modified-Wald (+) and Wald (o) procedures. Both regions are really similar and the 2 symbols cannot really be differentiated.

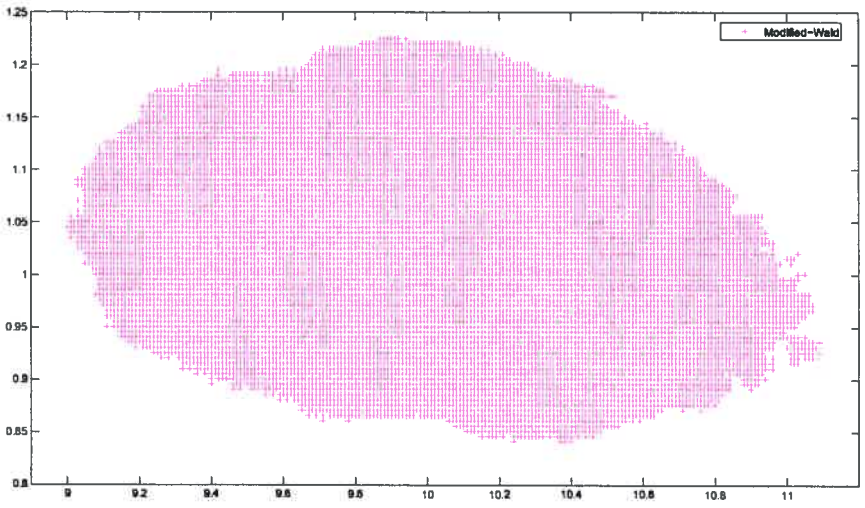

Figure III.2: 85%-Averaged Confidence Region when b=.1 using Modified-Wald (+). The ACR is empty for Wald.

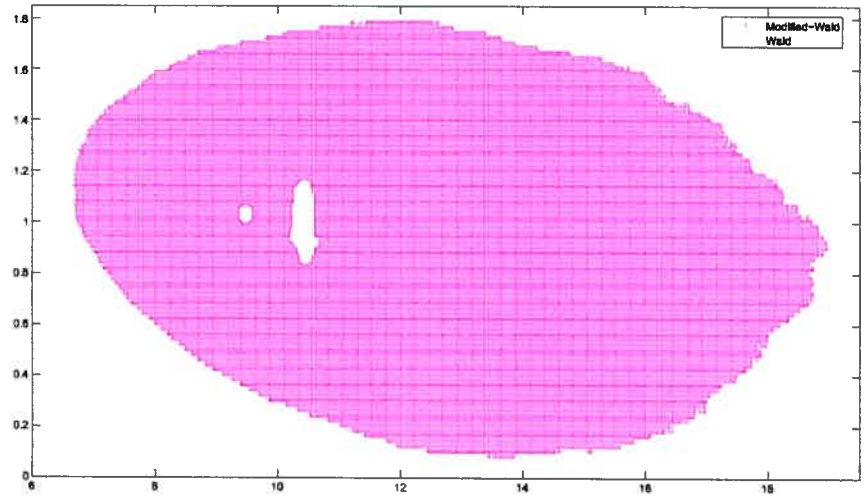

Figure III.3: 15%-Averaged Confidence Region when b=.1 using Modified-Wald and Wald procedures. The black area (including the white spots) represents the ACR for Modified-Wald. The white spots (inside the black area) represent the ACR for Wald.

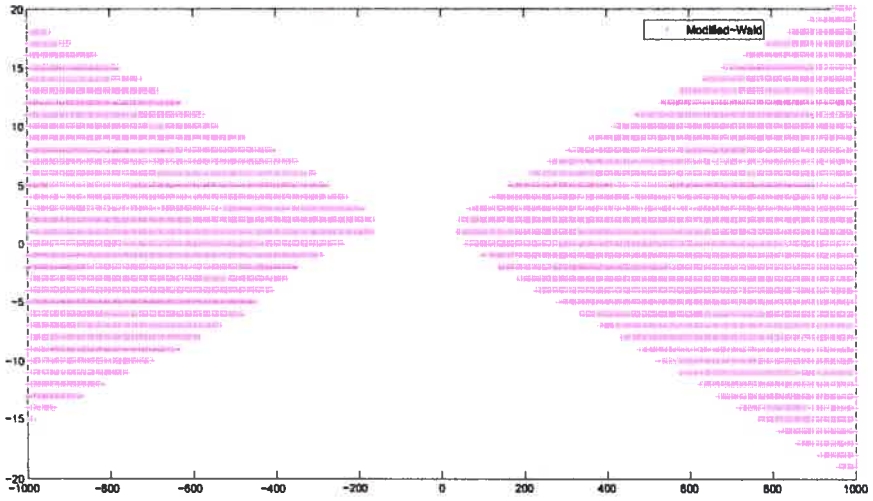

Figure III.4: 75%-Averaged Confidence Region when b=.01 using Modified-Wald (+). The ACR is empty for Wald.

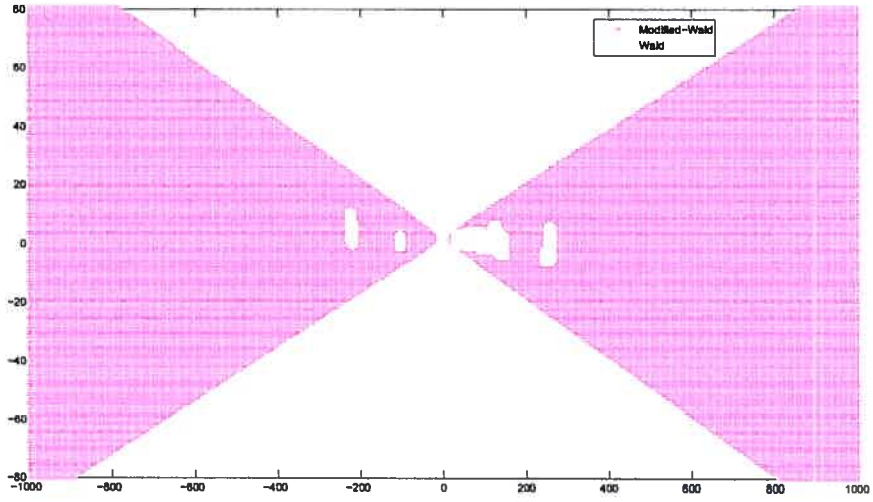

Figure III.5: 1%-Averaged Confidence Region when b=.01 using Modified-Wald and Wald procedures. The black area (including the white spots) represents the ACR for Modified-Wald. The white spots (inside the black area) represent the ACR for Wald.

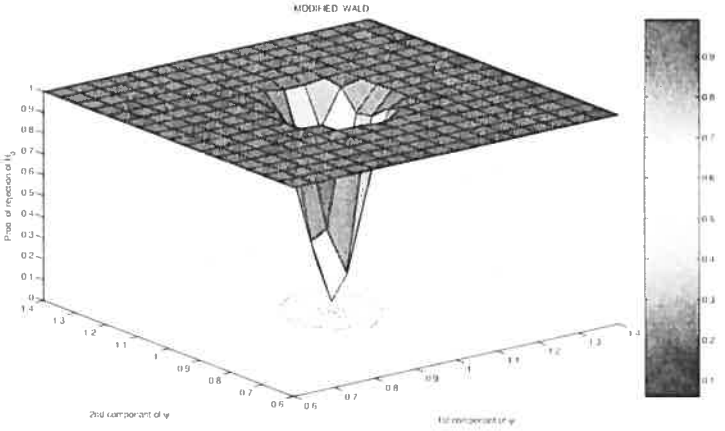

Figure III.6: Power function when b=1 using Modified-Wald procedure.

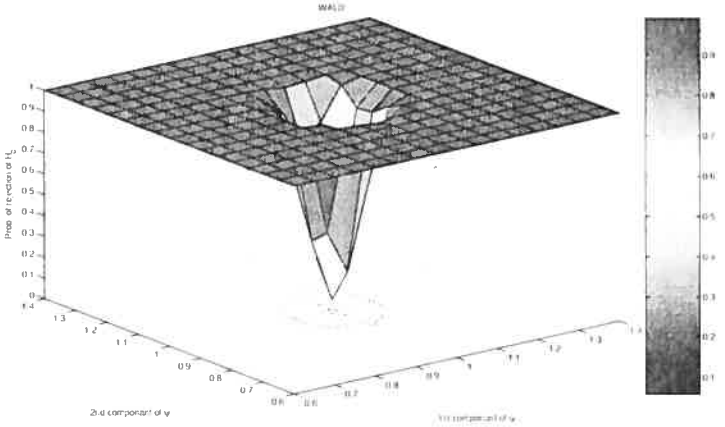

Figure III.7: Power function when b=1 using Wald-type procedure.

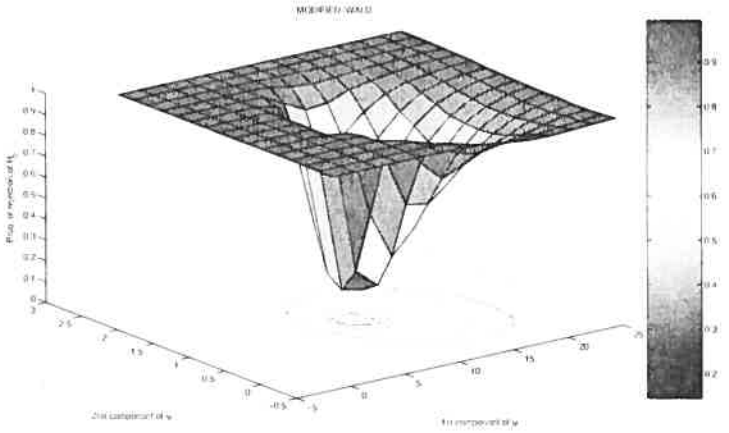

Figure III.8: Power function when b=.1 using Modified-Wald procedure.

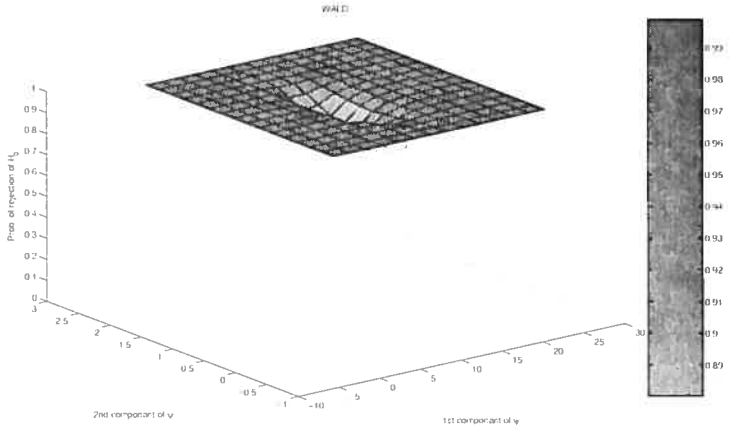

Figure III.9: Power function when b=.1 using Wald-type procedure.

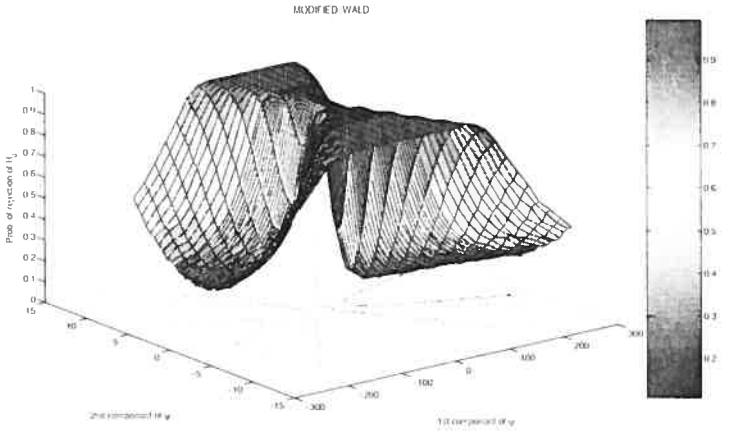

Figure III.10: Power function when b=.01 using Modified-Wald procedure.

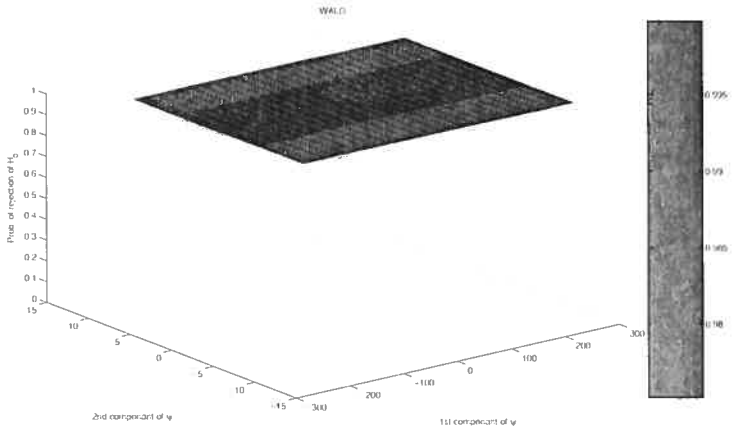

Figure III.11: Power function when b=.01 using Wald-type procedure.

Chapitre IV

Portfolio Selection with Estimation Risk: a Test Based Approach

1 Introduction

An optimal portfolio is the *best* allocation of funds across available assets[1]. Of course, what best means depends on the performance measure we use. Markowitz (1959) offers the classic definition of portfolio efficiency: a portfolio is efficient if it has the largest expected return for a given level of risk. For a given level of risk-aversion, this mean-variance efficiency provides a convenient single-period framework and remains among the most important benchmark models used by practitioners nowadays (see Meucci (2005)). In practice, however, its associated optimal investment rule depends on unknown parameters, the mean and the variance of returns distribution. To get a feasible version of this optimal rule, Markowitz (1959) simply replaces the unknown parameters by some estimates. Applying such a plug-in method gives rise to several issues. First the estimation risk is overlooked: in practice samples are finite, hence estimates are different from their respective true values. This new source of risk even appears in well-specified parametric models and adds to the traditional financial risk[2]. Second, is this feasible rule optimal? A (suboptimal) two-step approach can only be motivated when one believes that the estimated rule is *not too far* from the true optimal one.

In sharp contrast with existing literature, we focus on a different measure of performance. We borrow from practitioners and evaluate different funds allocations through their likelihood of beating a benchmark. Several industries are actually interested in such a goal: for instance, institutional money managers, and among others the defined benefits pension plans and the endowment plans are devoted to guarantee the (chosen) minimal performance. For a given benchmark, we deduce a closed-form and workable optimal investment rule which naturally incorporates the estimation risk of the mean, and does not depend on any nuisance parameter. Hence it is directly applicable without requiring any additional (suboptimal) plug-in step.

More precisely, our portfolio selection method is based on a one-sided test ensuring that the portfolio performance is above a given threshold; then we obtain the optimal allocation from the maximization of the associated p-value. The specific design of the p-value selection method has three advantages. First, the test is a natural and valid statistical tool to compare random quantities (here the estimated portfolio performance). Hence the uncertainty of the problem is directly accounted for: we will see that this is crucial to get an exact optimal

[1] Brandt (2004) provides a broad survey on general issues related to portfolio choice.

[2] Kan and Zhou (2006) provide an extensive study of the financial consequences of ignoring estimation risk.

investment rule. Second, maximizing the p-value increases the likelihood of our objective of interest (here to beat the chosen benchmark). Finally the optimal investment rule belongs to the class of two-fund investment rules, similar to the (feasible) optimal mean-variance rule: investing in the (sample) tangency portfolio and in the riskless asset[3]. This investment rule corresponds to a mean-variance investor with a corrected, sample-dependent risk-aversion parameter. While existing literature recommends to increase the risk-aversion parameter to account for estimation risk, we advocate more flexibility: we may indeed decrease the risk-aversion parameter depending on the realized sample.

Our work relates to the literature as follows. First, estimation risk in portfolio allocation has been known for a while. One of the earliest and maybe most natural solution appears to be Bayesian. The Bayesian approach is based on the predictive distribution introduced by Zellner and Chetty (1965) under which expectations are now considered. It provides a general framework where estimation risk is naturally accounted for when considering the parameters as random variables: the posterior distribution captures their possible outcomes and is combined to a prior model to derive the predictive distribution. The study by Bawa, Brown and Klein (1979) surveys the early literature, and is then followed by many others. It is not always clear how the prior model can be chosen, even though it is based on the investor's knowledge and experience: different priors may lead to contrastive investment strategies. We only consider non-informative prior models.

More recently, some authors decided to directly focus on the expected financial loss when the optimal investment rule is replaced by some feasible one. Due to the complexity of the problem, ter Horst, de Roon and Werker (2006) and Kan and Zhou (2006) restrict their attention to the class of two-fund investment rules (similar to the feasible optimal mean-variance rule). While ter Horst *et al.* (2006) ignore the estimation risk of the variance, Kan and Khou (2006) (under the normality assumption of the returns) provide a closed-form solution to the simplified problem. However, the optimal rule depends on nuisance parameters. So, in order to implement this optimal strategy, one needs to add a suboptimal plug-in step[4]. The mean-

[3]This specific class of investment rules has already been considered in the literature: see ter Horst, de Roon and Werker (2006) and Kan and Zhou (2006). However, here it directly follows from our portfolio selection method (just like the mean-variance procedure) and not from a simplifying assumption.

[4]Of course, by construction, Kan and Zhou (2006) theoretical two-fund rule outperforms any p-value investment rule. However nothing is guaranteed when one considers its feasible version as shown in our simulation exercise.

variance framework seems to be limited. As shown by Kan and Zhou (2006), who were able to exhibit the optimal two-fund rule while accounting for estimation risk, the outcome is not completely satisfactory as the optimal rule is unfeasible. This more general issue arises when one maximizes some expected quantity. This motivates our approach: we take some distance with the traditional minimization of an expected financial loss function and maximize the likelihood of some desirable event[5].

Finally, previous studies have already focused on defeating a benchmark: see for instance Stutzer (2003) and references therein. However, to our knowledge, this has not yet been related to estimation risk. Moreover, these studies work in a continuous time framework: this is definitely not our interest here.

To conclude, a simple Monte-Carlo study involving five risky assets (calibrated from monthly unhedged returns of stock indices for the G5 countries) is used to compare eleven investment strategies. These are compared with respect to their out-of-sample expected performances as well as with respect to their maintenance costs and stability over some investment horizon. The p-value selection method performs surprisingly well considering it is not specifically designed to maximize the mean-variance performance. Moreover, it avoids extreme positions in the assets and remains relatively stable over time.

The remainder of the chapter is organized as follows. Section 2 solves the classical mean-variance problem. The p-value selection method is introduced in section 3. Section 4 reviews some competing investment strategies. Section 5 presents the results of a simple simulation study calibrated from real data. Section 6 concludes.
The details of the calculations are gathered in the appendix.

2 Classical Mean-Variance problem

This section discusses the mean-variance problem and introduces estimation risk. Consider an investor who chooses a portfolio among N financial risky assets and the riskless asset. At

[5]Others have also departed from the classical mean-variance approach: Garlappi, Uppal and Wang (2006) propose a sequential max-min method where the worst performance (when the unknown parameters fall into a confidence interval) is maximized with respect to the portfolio weights; Harvey, Liechty, Liechty and Muller (2004) adopt a Bayesian setting under the assumption that the returns follow a skew-normal distribution.

time t, denote respectively by $R_t \equiv (r_{1t} \cdots r_{Nt})'$ and R_{ft} the rates of returns on the N risky assets and the riskless asset. The vector of excess returns is defined as $\tilde{R}_t \equiv R_t - R_{ft}\iota$ where ι is the conformable vector of ones. The following standard assumption is maintained on the probability distribution of excess returns $\tilde{R}_t$:

Assumption 1. *The vector of excess returns $\tilde{R}_t$ is independent and identically distributed over time. In addition, R_t is normally distributed with mean $\tilde{\mu}_0$ and variance Σ_0.*

At time t, the portfolio is built after investing a vector θ into the risky assets and $(1 - \theta'\iota)$ in the riskless asset. The portfolio excess return is $r_t^P(\theta) \equiv \theta'\tilde{R}_t$, and its associated mean and variance are then respectively,

$$\mu_P = \theta'\tilde{\mu}_0 \quad \text{and} \quad \sigma_P^2 = \theta'\Sigma_0\theta$$

The vector of weights θ defines the investment rule which maximizes the following mean-variance objective function:

$$\max_\theta \left\{ E\left[r_t^P(\theta)\right] - \frac{\eta}{2} Var\left[r_t^P(\theta)\right] \right\} \iff \max_\theta \left\{ \theta'\tilde{\mu}_0 - \frac{\eta}{2}\theta'\Sigma_0\theta \right\}$$

where η is the coefficient of relative risk-aversion. This leads to the following optimal vector of weights:

$$\theta_{MV} = \frac{1}{\eta}\Sigma_0^{-1}\tilde{\mu}_0 \tag{2.1}$$

In practice, the parameters $\tilde{\mu}_0$ and Σ_0 are unknown: therefore the optimal mean-variance investment rule θ_{MV} is *unfeasible* and cannot be calculated in practice. Markowitz (1959) simply replaces the unknown parameters by some estimates. This easily provides a *feasible* version of the above optimal rule. More precisely, for some estimates $\hat{\tilde{\mu}}$ and $\hat{\Sigma}$ of the unknown parameters $\tilde{\mu}_0$ and Σ_0, one defines the feasible (random) investment rule and its associated (random) performance as:

$$\hat{\theta}_{MV} = \frac{1}{\eta}\hat{\Sigma}^{-1}\hat{\tilde{\mu}} \quad \text{and} \quad Q(\hat{\theta}_{MV}) = \frac{1}{2\eta}\hat{\tilde{\mu}}'\hat{\Sigma}^{-1}\hat{\tilde{\mu}} \tag{2.2}$$

where $\hat{\tilde{\mu}}$ and $\hat{\Sigma}$ are, for instance, the maximum likelihood estimators,

$$\hat{\tilde{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{R}_t \quad \text{and} \quad \hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(\tilde{R}_t - \hat{\tilde{\mu}})(\tilde{R}_t - \hat{\tilde{\mu}})' \tag{2.3}$$

Applying this *plug-in* method comes at a price. First, estimation risk is overlooked. In practice, the sample size is only T (finite), hence $\hat{\mu}$ and $\hat{\Sigma}$ are different from their respective true values. Second, precisely because the feasible rule $\hat{\theta}_{MV}$ is numerically different from the true optimal one, its optimality cannot be guaranteed. In the next section, we propose a portfolio selection method that incorporates estimation risk and does not require any additional (suboptimal) step.

3 Maximization of the p-value

This section introduces the p-value selection method and derives the associated optimal investment rule for a given benchmark c. In a second step, the question of the existence of an optimal benchmark is raised.

3.1 Definition and Optimal investment rule

As emphasized earlier, this chapter takes some distance with the classical mean-variance framework and the common idea of minimizing some (expected) financial risk function. More precisely, in sharp contrast with existing literature, we do not maximize any usual measure of portfolio performance. We rather compare available funds allocations through their likelihood of beating the chosen benchmark. Of course, our portfolio selection method crucially depends on the benchmark. *Reasonable* benchmark choices yield to more *conservative* objective functions than the classic maximization of the (mean-variance) performance. Our investor is more conservative in the sense that she is not interested in achieving the maximal performance at every period; she rather selects the investment rule that maximizes the likelihood of defeating the benchmark. This selection method directly accounts for the random nature of the problem while being of primary concern for several industries, like institutional money managers.

Our portfolio selection method is based on a one-sided test that the chosen measure of portfolio performance is above the given threshold. Obviously, two unknowns remain here: first the choice of the performance measure and second the threshold. As pointed out earlier, Markowitz's mean-variance efficiency is a convenient framework privileged by practitioners.

Accordingly, we consider the following measure of portfolio performance:

$$Q(\mu_P, \sigma_P^2) = \mu_P - \frac{\eta}{2}\sigma_P^2 \tag{3.1}$$

where (μ_P, σ_P^2) are respectively the first two moments of the probability distribution of the portfolio. This measure of performance has mainly been chosen for comparison purposes: our p-value selection method works with any other measure $Q(.)$[6]. Not only the test is the natural statistical tool to compare random quantities and incorporate estimation risk, but also it directly focuses on the well-defined objective for a portfolio manager, to beat the performance of a benchmark index.

Formally, the null hypothesis of interest is stated as follows:

$$H_0 : \ Q(\mu_P, \sigma_P^2) > c \tag{3.2}$$

where c is the (deterministic) performance of the (chosen) benchmark index. To construct the associated test statistic, some assumptions are needed on the probability distribution of the returns. Consider an investor at time T who has observed the N risky asset returns from time $t = 1$ to T.

Assumption 2. *The vectors of the N financial excess returns of interest at time t, $\tilde{R}_t = [\tilde{r}_{1t} \cdots \tilde{r}_{Nt}]'$ for t=1 to T, are identically distributed and serially independent. More formally,*

1) $\tilde{R}_t \sim \mathcal{F}(\tilde{\mu}_0, \Sigma_0) \ \forall \ t = 1 \cdots T$ *where $\mathcal{F}$ is some smooth distribution function whose first two moments exist*
2) $\tilde{R}_t$ *and* $\tilde{R}_{t'}$ *are independent* $\forall \ (t, t') \in \{1, 2, \cdots, T\} \ / \ t \neq t'$

We consider from now on the portfolio excess return $\tilde{r}_t^P(\theta) = \theta' \tilde{R}_t$. Note that this only shifts the deterministic benchmark c: only strictly positive benchmarks c are now considered. A null benchmark corresponds to the minimal acceptable performance, guaranteed when always investing in the riskless asset. The measure of portfolio performance is then written as:

$$Q_P(\theta) = E\tilde{r}_t^P(\theta) - \frac{\eta}{2}Var(\tilde{r}_t^P(\theta))$$

[6] Any performance measure works under minor assumption like assumption 2. In particular, we could think of incorporating higher moments to account for effects of skewness, kurtosis... This only affects the tractability of the optimal investment rule when one wants to account for the associated estimation risk. This is indeed related to the (asymptotic) distribution of the estimated portfolio performance (that may need to be simulated).

and is estimated by[7]:

$$\hat{Q}_P(\theta) = \theta'\hat{\tilde{\mu}} - \frac{\eta}{2}\theta'\hat{\Sigma}\theta \tag{3.3}$$

$$\text{with } \hat{\tilde{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{R}_t \text{ and } \hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(\tilde{R}_t - \hat{\tilde{\mu}})(\tilde{R}_t - \hat{\tilde{\mu}})' \tag{3.4}$$

The application of the vectorial central limit theorem yields the asymptotic distribution of the estimated performance: $\sqrt{T}\left[\hat{Q}_P(\theta) - Q_P(\theta)\right]$ is asymptotically normally distributed with mean 0 and variance $Var(\hat{Q}_P(\theta))$. Then, for an estimator $\hat{S}$ of its standard deviation, the test statistic and associated p-value are defined as follows:

$$St(\theta) = \frac{\hat{Q}_P(\theta) - c}{\hat{S}/\sqrt{T}} \text{ and p-value}(\theta) = \int_{-\infty}^{St(\theta)} f_T(u)du$$

with f_T the density function of a student random variable with $(T-1)$ degrees of freedom. Hence the maximization problem is finally stated as:

$$\max_{\theta}\left[\text{p-value}(\theta)\right] \iff \max_{\theta}\left[St(\theta)\right]$$

The p-value selection method can be linked to the well-known financial risk measure, the Value-at-Risk (VaR hereafter). Briefly the VaR at level α represents an estimate of the level of loss on a portfolio which is expected to be equaled or exceeded with the given, small probability α: risk regulations usually dictates the choice of this level of confidence. Our selection method rather guarantees the chosen minimal level of performance with the highest level of confidence. We think that choosing the benchmark is more inline with institutional money managers concerns. See the appendix for an extended discussion.

Obviously, estimation risk is related to the estimation of both the mean and the variance of the portfolio. If it is commonly accepted that the estimation error on the sample mean is much larger than on the sample variance, recent studies suggest that it might not always be the case: see e.g. Cho (2003) and Kan and Zhou (2006). The latter authors conclude that this is only acceptable when N/T is *small*: in particular there is an interactive effect between both estimation errors. Here, to simplify the problem (and get an interpretable closed-form investment rule), we ignore the estimation risk of the variance[8]. The simplified maximization

[7] The procedure remains similar for any other set of consistent estimates. We could even think of the selection problem as starting right here, with a set of estimates given by a practitioner.

[8] Note that, in our simulation study in section 5, the ratio N/T is kept small.

problem is now:

$$\theta_p(c) = \arg\max_{\theta} \left[\frac{\theta'\hat{\tilde{\mu}} - (\eta/2)(\theta'\hat{\Sigma}\theta) - c}{(\theta'\hat{\Sigma}\theta)^{1/2}/\sqrt{T}} \right]$$

where $\hat{\tilde{\mu}}$ and $\hat{\Sigma}$ have been defined in equation (3.4).

Definition 3.1. *Let $\hat{\tilde{\mu}}$ and $\hat{\Sigma}$ respectively be estimators of the first two moments of the distribution of the excess returns as in (3.4). Then, for a given (deterministic) benchmark c, the optimal p-value investment rule is defined as:*

$$\theta_p(c) = \sqrt{\frac{2\eta c}{\hat{\tilde{\mu}}'\hat{\Sigma}^{-1}\hat{\tilde{\mu}}}} \frac{1}{\eta} \hat{\Sigma}^{-1}\hat{\tilde{\mu}} \tag{3.5}$$

Several comments are worth mentioning. First, the optimal p-value rule $\theta_p(c)$ is random and depends on the (chosen) estimates of the mean and variance of the excess returns distribution. However, this random rule (3.5) is the genuine rule that solves our optimization problem. In other words, our workable rule does not come from an additional (suboptimal) plug-in step (see also section 4). The deep reason for this exactness lies in the definition of our p-value selection method: the randomness of the problem precisely defines our selection procedure. Without uncertainty, there would not be any purpose to run a test and therefore no p-value maximization. Second, the rule (3.5) is a two-fund investment rule, just like the (feasible) mean-variance optimization problem $\hat{\theta}_{MV}$ (see equation (2.2)): both rules yield to the same repartition of wealth among the different financial risky assets. This allows us to reinterpret the p-value investor in terms of mean-variance behavior with a corrected risk-aversion parameter in section 4. Finally, note that the optimal p-value investment rule works for a given c. The next section naturally asks whether there exists an optimal benchmark or not.

3.2 An optimal choice for the benchmark?

The above selection method depends on the choice of the benchmark c: it represents the minimal level of portfolio performance the investor wants to guarantee with the highest possible level of confidence. In some sense, this benchmark is not a choice variable and we cannot really talk about its optimality. However, it is helpful to exhibit the *optimal* benchmark for comparison purposes.

We maximize the expected performance of the portfolio associated with the optimal p-value rule for a given benchmark:

$$\max_{c>0} E\left[Q_P(\theta_p(c))\right]$$

The *optimal* benchmark $c^\star$ reads:

$$c^\star = \frac{1}{2\eta} \times \frac{\left[E\left(\frac{\tilde{\mu}'\Sigma^{-1}\tilde{\mu}_0}{\sqrt{\tilde{\gamma}^2}}\right)\right]^2}{\left[E\left(\frac{\tilde{\mu}'\Sigma^{-1}\Sigma_0\Sigma^{-1}\tilde{\mu}}{\tilde{\gamma}^2}\right)\right]^2} \quad \text{where } \tilde{\gamma}^2 \equiv \tilde{\mu}'\Sigma^{-1}\tilde{\mu} \tag{3.6}$$

The optimal benchmark $c^\star$ is clearly unfeasible since it depends on the unknown parameters $\tilde{\mu}_0$ and Σ_0[9]. Interestingly, without estimation risk (or assuming we know $\tilde{\mu}_0$ and Σ_0), we can check that the associated investment rule is numerically equal to the true mean-variance rule, which is also the optimal rule in absence of estimation risk. See also section 4.3.

4 Theoretical comparison with existing literature

This section is dedicated to the comparison of competing investment strategies after introducing the useful concept of *corrected* risk-aversion parameter, already considered in ter Horst *et al.* (2006).

4.1 Overview of some competing selection methods

This subsection briefly introduces some of the existing investment rules. In particular, we emphasize the different methodologies to account for estimation risk[10].

• Mean-variance (Markowitz (1959)) (see section 2): this rule selects the portfolio with the maximal mean-variance performance. The optimal allocation is unfeasible: it depends on the first two unknown moments of the excess returns distribution. A simple variation, where some estimates (see equation (2.3)) of the unknowns are plugged into the formula, becomes feasible. The estimation risk is then ignored. This rule is given by:

$$\tilde{\theta}_{MV^\star} = \frac{1}{\eta}\Sigma^{-1}\tilde{\mu}$$

[9] This is not really surprising since we maximize the expected performance for a given c.

[10] See also Kan and Zhou (2006).

• Bayesian (Bawa, Brown and Klein (1979)): the Bayesian approach maximizes the expected performance of the portfolio where the expectation is computed according to the predictive distribution of the market. In turn, this predictive distribution is built from a combination of historical observations and the prior. Estimation risk is made explicit by considering the unknown parameters as random variables, described by the posterior distribution. However, it is not always clear how the prior can be chosen. Under the standard assumption of diffuse priors on both the mean and the variance of the excess returns, it can be shown that the Bayesian optimal portfolio weights are:

$$\theta_B = \frac{1}{\eta}\left(\frac{T-N-2}{T+1}\right)\hat{\Sigma}^{-1}\hat{\tilde{\mu}}$$

• ter Horst, de Roon and Werker (2006): the portfolio weights are chosen to minimize the risk function based on the loss of replacing the true (unknown) mean of the portfolio by its sample estimate. They restrict their attention to the class of two-fund rules and ignore the estimation risk of the variance:

$$\theta_{HRW} = \frac{1}{\eta}\left(\frac{\gamma^2}{\gamma^2 + N/T}\right)\hat{\Sigma}^{-1}\hat{\tilde{\mu}} \quad \text{with} \quad \gamma^2 = \tilde{\mu}_0'\Sigma_0^{-1}\tilde{\mu}_0$$

The resulting optimal rule θ_{HRW} is unfeasible: γ^2 is then replaced by its sample counterpart $\hat{\gamma}^2 = \hat{\tilde{\mu}}\hat{\Sigma}^{-1}\hat{\tilde{\mu}}$. Again, optimality is not guaranteed.

• Kan and Zhou (2006): they extend the previous selection method to incorporate the estimation risk of the variance:

$$\theta_{KZ2} = \frac{1}{\eta}\left(\frac{(T-N-4)(T-N-1)}{T(T-2)} \times \frac{\hat{\gamma}^2}{\hat{\gamma}^2 + N/T}\right)\hat{\Sigma}^{-1}\hat{\tilde{\mu}}$$

Just like θ_{HRW}, the resulting optimal rule θ_{KZ2} is unfeasible: see appendix B for its feasible version. They also explore the class of three-fund investment rules by considering the sample global mean-variance portfolio. The associated optimal rule θ_{KZ3} is unfeasible as well: see also appendix B for additional details.

• Garlappi, Uppal and Wang (2006): they consider a model that allows for multi priors and where the investor is averse to ambiguity. The standard mean-variance framework is modified by adding a preliminary minimization step. A constraint restricts the expected return to fall into a confidence interval around its estimated value and recognizes the

existence of estimation risk. The minimization over the possible expected returns subject to this constraint reflects the investor's aversion to ambiguity. While this approach has a solid axiomatic foundation, its sequentiality cannot be directly liked to an optimality criterion. The optimal rule θ_{GIW} is defined in appendix B.

The following theoretical rankings have been derived by Kan and Zhou (2006):

$$\theta_{KZ3} >> \theta_{KZ2} >> \theta_B >> \theta_{MV}, \quad \theta_{KZ2} >> \theta_{HRW} \quad \text{and} \quad \theta_{KZ2} >> \theta_{GIW}$$

where $>>$ stands for "outperforms in terms of mean-variance performance". We argue that this ranking might not be guaranteed in practice (even in simple simulation frameworks where the returns are normally distributed) when θ_{HRW}, θ_{KZ2} and θ_{KZ3} are replaced by their feasible counterparts. Kan and Zhou (2006) already mentioned this issue when comparing their (feasible) optimal two-fund rule to the one of Garlappi *et al.*. See also section 5.

4.2 Corrected risk-aversion parameter

Despite their differences, most of the selection methods described above yield an optimal rule within the class of two-fund rules, just like the (feasible) Markowitz's mean-variance approach[11].

According to these rules, the same repartition of wealth among the different risky financial assets is recommended: their differences lie in the share of wealth invested in risky assets relative to the riskless asset. The (feasible) mean-variance rule can be reinterpreted as a function of the risk-aversion parameter η:

$$\hat{\theta}_{MV}(\eta) = \frac{1}{\eta}\hat{\Sigma}^{-1}\hat{\mu}$$

Note that $\left[\hat{\Sigma}^{-1}\hat{\mu}\right]$ defines how wealth is allocated among risky assets while η weights the share of wealth assigned to the risky assets: the greater η, the lower the (global) share to the risky assets.

[11] This is especially surprising for our p-value selection method since it does not come from any simplifying assumption (as for θ_{KZ2} and θ_{HRW}).

We can then write each two-fund rule as a mean-variance rule with a *corrected* risk aversion parameter. In fact, any two-fund rule vector of weights θ_r can be rewritten as follows:

$$\theta_r = \theta_{MV}(\tilde{\eta}) \quad \text{for some} \quad \tilde{\eta} > 0 \tag{4.1}$$

Therefore, the behavior of any two-fund investor can be characterized in terms of a mean-variance associated to a new (corrected) risk-aversion parameter $\tilde{\eta}$. The following corrected risk-aversion parameters can be deduced for the two-fund rules discussed above[12]:

$$\begin{aligned}
\tilde{\eta}_{HRW} &= \eta \frac{\gamma^2 + N/T}{\gamma^2} \\
\tilde{\eta}_{KZ2} &= \eta \frac{\gamma^2 + N/T}{\gamma^2} \times \frac{T(T-2)}{(T-N-4)(T-N-1)} \\
\tilde{\eta}_B &= \eta \frac{T+1}{T-N-2} \\
\tilde{\eta}_p(c) &= \eta \sqrt{\frac{\tilde{\mu}' \hat{\Sigma}^{-1} \tilde{\mu}}{2\eta c}}
\end{aligned}$$

4.3 Comparison of the reinterpreted investment rules

Our original mean-variance investor always becomes more risk-averse when applying any of the competing rules we consider here. However, this is not true when she applies the p-value rules: her risk-aversion parameter does not always increase.

On one hand, the investors respectively associated with the three competing rules θ_B, θ_{HRW} and θ_{KZ2} are always more risk-averse than the mean-variance investor. Moreover the following ranking can even be observed:

$$\tilde{\eta}_{KZ2} > \tilde{\eta}_{HRW} > \eta \quad \text{and} \quad \tilde{\eta}_B > \eta$$

Recall that θ_{KZ2} is nothing but θ_{HRW} where the additional estimation risk coming from the variance is accounted for. So one could be tempted to conclude that increasing the risk-aversion parameter is a sensible way to account for estimation risk.

[12] We could also consider θ_{GIRW} as a two-fund rule with a corrected risk-aversion parameter that can be infinite with non-zero probability.

On the other hand, the p-value corrected risk-aversion linearly depends on the original risk-aversion parameter: hence, the p-value investor might be characterized as a mean-variance investor either by increasing or decreasing the risk-aversion η. The corrected risk-aversion parameter can actually be rewritten as follows:

$$\tilde{\eta}_p(c) = \eta \sqrt{\frac{Q(\hat{\theta}_{MV})}{c}}$$

where $Q(\hat{\theta}_{MV})$ is the performance associated to the feasible mean-variance investment rule (see equation (2.2)). Depending on the choice of the benchmark c, one falls into one of the following cases:

(i) if $c = Q(\hat{\theta}_{MV})$ then $\tilde{\eta}_p = \eta$
(ii) if $c > Q(\hat{\theta}_{MV})$ then $\tilde{\eta}_p < \eta$
(iii) if $c < Q(\hat{\theta}_{MV})$ then $\tilde{\eta}_p > \eta$

Intuitively, this additional flexibility might be profitable, especially because it can be linked to the actual sample realizations. Consider an investor who chooses a moderate benchmark c. Assume now that, *by chance*, she faces a good financial environment (or a sample associated to a relatively high performance): likely $c < Q(\hat{\theta}_{MV})$ and so $\tilde{\eta}_p > \eta$. Overall, the part invested in the risky assets is going to be lower. The profitable financial conditions offer additional safety to the p-value investor: it is more likely to beat the target. On the contrary, with a *not so good* financial environment, one may expect the investor to become less risk-averse, still hoping to defeat the benchmark. Intuitively, it makes sense to incorporate the sample-information into the decision process. The p-value selection method might also overcome the well-known problem of the mean-variance investment rule which takes extreme positions. The next section further investigates this.

5 Monte-Carlo study

This section presents the results of a simple Monte-Carlo study. The simulation exercise involves five risky assets and the riskless asset. The risky returns follow a multivariate normal distribution and the true model parameters are calibrated from monthly unhedged returns of stock indices for the G5 countries over the period January 1974 to December 1998. The G5

stock indices are the MSCI indices for France, Germany, Japan, the UK and the US as done, for instance, in ter Horst *et al.* (2006). Table IV.1 contains the summary statistics.

A financial strategy is considered over an investment horizon T_h. More precisely, at time $t = 1$ investors have access to T (past) historical observations of the financial returns. These are used to estimate the unknown parameters (typically $\hat{\mu}_0$ and Σ_0) required to evaluate their investment strategy. The induced portfolio is hold for one period until $t = 2$, whereas the investment strategy is reevaluated using again the T most recent observations to build the estimators. A new portfolio is constructed, and so on until T_h.

We compare eleven investment strategies around two objects of interest for portfolio managers. First, we compare their respective performance over the investment horizon. The performance is evaluated through the expected (one-period) mean-variance performance. Second, we compare the stability of the investment rules as measured by the transaction costs incurred to reallocate the portfolio at each period.

We consider the following investment rules: (1) the mean-variance optimal rule in absence of uncertainty θ_{MV}; (1f) the feasible counterpart of (1) $\hat{\theta}_{MV}$; (2) the optimal two-fund rule θ_{KZ2}; (2f) the feasible counterpart of (2) $\hat{\theta}_{KZ2}$; (3) the optimal two-fund rule when the variance is known θ_{HRW}; (3f) the feasible counterpart of (3) $\hat{\theta}_{HRW}$; (4) the Bayesian rule with diffuse prior θ_B; (5) the sequential min-max rule θ_{GRW}; (6) the optimal three-fund rule θ_{KZ3}; (6f) the feasible counterpart of (6) $\hat{\theta}_{KZ3}$; (7) the p-value rule for four different benchmarks. In this convenient Monte-Carlo setup, the benchmarks can be evaluated directly with respect to the maximal performance $Q(\theta_{MV})$. Typically, we consider here $c_1 = .1Q(\theta_{MV})$; $c_2 = .5Q(\theta_{MV})$; $c_3 = .9Q(\theta_{MV})$; and c^* the *optimal* benchmark (according to section 3.2) which is evaluated by simulation for each size of the rolling window (see table IV.2). In practice, one can think of at least two ways to get a convenient benchmark: c might be a numerical target that has been decided by the board of directors; c can also be based on the historical performance of some benchmark index.

We choose to set the risk-aversion parameter η equal to 5. For each portfolio rule r, defined by the vector of weights θ_r, the associated (one-period) expected performance is evaluated as follows[13]:

$$E\left[Q(\theta_r)\right] = E\left(\theta_r'\hat{\mu}_0\right) - \frac{\eta}{2}E\left(\theta_r'\Sigma_0\theta_r\right) \tag{5.1}$$

[13]Note that to simplify the notations we do not make explicit its dependence to the date of the investment

where the true moments μ_0 and Σ_0 are known (but only at this stage!) in our convenient Monte-Carlo framework: this helps isolate the effects of estimation risk.

Most of the above rules lead to a random vector of weights θ_r. Hence, this is not always possible to obtain a closed-form solution for the expected performance. If so, the performance is evaluated by averaging over many replications of the experiment. This is the case for the rules (2f), (3f), (5), (6f) and (7). For the remaining rules, expected performances are formally provided in appendix B.

Table IV.3 provides the expected performances (in percentages per month) associated with every rule for several sample sizes of the rolling window used to calculate the estimators.
Generally speaking, things get better when the sample size increases: i) the performance of each investment rule gets closer to the true optimal one $Q(\theta_{MV})$; ii) the feasible rules get closer to their theoretical counterparts - see also table IV.4; iii) finally, the estimation risk coming from the variance matters less when T increases. There is an additional loss of 15% per month when using $\hat{\theta}_{HRW}$ instead of $\hat{\theta}_{KZ2}$ for $T = 120$ and it drops to less than 1% when $T = 240$.
Figure IV.1 provides a visual comparison of the performances of all the feasible rules, as a function of the rolling window size. The dominance of the feasible three-fund rule is obvious. Hence, diversification appears to matter quite a bit when accounting for estimation risk. The p-value with a medium benchmark performs fairly well.
Figure IV.2 provides the same information for the p-value rules compared to their associated target. Note first that the rank of the expected performances is preserved: a low target is associated to a lower expected performance. Then, except for the highest target (chosen as 90% of the maximal theoretical performance), the minimal target is always ensured and indeed outperformed.
The performance of the p-value investment rule is positively surprising. Recall that compared to most of its competitors, it does not maximize the mean-variance performance. Of course, its performance crucially depends on the benchmark[14]. However, for quite a wide range of potential benchmarks ($c_1 = .035$ to $c_3 = .315$ percentages per month) the p-value performs quite well: the medium benchmark even outperforms $\hat{\theta}_{KZ3}$ when $T = 60$. It clearly outperforms the mean-variance, the Bayesian and the min-max sequential investment rules.

We now compare the stability of the portfolio rules via the transaction costs incurred to real-

[14]Simulations with *unreasonable* targets, both very low and high, confirm this. They are not reported here.

locate the portfolio at each period. This cost is the averaged amount (in arbitrary units) payed by the investor to modify her positions. The arbitrary cost is the same for each risky asset. More precisely, for each rule r defined by the vector of weights θ_t, at date t, the maintenance cost is defined as follows:

$$C_r = E\left[\sum_{t=1}^{T_h-1} |\theta_{t+1} - \theta_t| \iota\right] \sim \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{t=1}^{T_h-1} |\theta_{t+1}^m - \theta_t^m| \iota\right] \tag{5.2}$$

where $\iota = [1\ 1\ \cdots\ 1]'$ and M is the number of replications.

Table IV.5 collects the average transaction costs for each rule, over an investment horizon $T_h = 60$ with several rolling window sizes and $M = 50000$ replications. Even though the transaction costs are calculated in (5.2) in a relatively basic and crude way, several comments are worth mentioning.

First, generally speaking, the transaction costs decrease when the size of the rolling window increases: the estimators naturally become more accurate and stable when the sample size increases, so do the financial positions. Only the most economical rule (θ_{GUW}) does not satisfy this. The reason comes from its definition. Contrary to any other investment rule, θ_{GUW} has a nonzero probability of entirely investing in the riskfree asset: this mechanically lowers its maintenance cost. Note that when the sample increases (and hence the estimators can be trusted more), the GUW-investor has higher transaction costs, meaning that she invests more in the risky assets.

The true (unfeasible) rules (θ_{KZ2}, θ_{HRW}, and θ_{KZ3}) tend to be more economical than their associated feasible counterparts. When the sample size increases, these rules, as well as most of the remaining ones, get closer to each others; this has already been noticed with the expected performance.

The p-value rules are naturally ranked as a function of their associated benchmark. More precisely, the lowest target c_1 yields a more economical rule $\theta_p(c_1)$. In order to fulfill her objective, this investor does not need to invest as much in the risky assets.

Finally the feasible rules can be ranked from the most economical as follows (the ranking does not depend on the size of the sample used to produce the estimators):

$$\theta_{GUW} >> \theta_p(c_1) >> \theta_p(c_2) \sim \hat{\theta}_{KZ3} >> \hat{\theta}_{KZ2} >> \theta_p(c_3) \sim \hat{\theta}_{HRW} >> \theta_B >> \hat{\theta}_{MV}$$

where $>>$ stands for "more economical" and $\sim$ for "economically equivalent".

6 Conclusion

In this chapter we propose a new way to account for estimation risk when selecting the optimal portfolio. In sharp contrast with existing literature, the optimal portfolio is not defined as the one maximizing some mean-variance performance: we consider here a more conservative definition of optimality which focuses on guaranteeing some minimal performance. More precisely, our portfolio selection method is based on a one-sided test ensuring that the portfolio performance is above a given threshold. The optimal weights are then obtained from the maximization of the p-value associated to the above test. The test provides an integrated method to account for estimation risk. Moreover, after neglecting the estimation risk of the sample variance, it leads to a closed-form investment rule which can be used without requiring any additional (suboptimal) step.

Of course the performance of the p-value investment rule (which is not designed or meant to achieve the maximal performance) depends on the chosen benchmark c. However, as illustrated in our simple Monte-Carlo simulation study where we consider a wide range of benchmarks, the overall performance is quite satisfactory. In particular, it performs pretty well for relatively small samples (we believe mainly because it does not require an additional suboptimal plug-in step) and outperforms reasonable choices of targets. We find these preliminary results really encouraging.

The great advantage of the simple framework we consider here consists in providing closed-form optimal investment rules, interpretable in terms of mean-variance behavior. Compared to competing two-fund rules (e.g. Kan and Zhou (2006) and ter Horst, de Roon and Werker (2006)), we have shown that this is not always an increase of the original risk-aversion parameter that works to account for estimation risk.

For future research, several directions might be worth examining. First, one could extend our selection method to random targets. This would permit to track the performance of benchmark indices, rather than numbers that may not always be inline with the financial environment. Second, considering that we generally do better than the feasible optimal two-fund rule and the very good results of the feasible three-fund rule, we may wonder how the p-value selection method, adapted to consider three-fund investment rules, would perform: as suggested by Kan and Zhou (2006), even more than three assets may help. Finally, recent papers have considered the related issue of model uncertainty. In particular, Cavadini, Sbuelz

and Trojani (2001) extend the study of ter Horst, de Roon and Werker (2006) to incorporate model risk: they use robust inference methods *à la Huber*, or local deviations to the chosen initial distribution. Of course, the interpretability of the investment rules is likely the price to pay to consider these extensions.

Appendix

Analogy with the Value-at-Risk

The Value-at-Risk (VaR) is a well-known financial risk measure summarizing the worst expected loss the investor is ready to accept. More precisely, the choice of a level of confidence $(1-\alpha)$ is associated to an α-quantile or $VaR(\alpha)$. When X represents the financial return of interest assumed normally distributed with mean μ and variance σ^2, we have:

$$\begin{aligned} P(X < -VaR_\alpha) = \alpha &\Leftrightarrow P\left(\frac{X-\mu}{\sigma} < \frac{-VaR_\alpha - \mu}{\sigma}\right) = \alpha \\ &\Leftrightarrow \Phi\left(\frac{-VaR_\alpha - \mu}{\sigma}\right) = \alpha \end{aligned}$$

where $\Phi(.)$ is the cumulative distribution function of a standard gaussian random variable. So,

$$\begin{aligned} P(X < -VaR_\alpha) = \alpha &\Leftrightarrow \frac{-VaR_\alpha - \mu}{\sigma} = \Phi^{-1}(\alpha) \\ &\Leftrightarrow -VaR_\alpha = \mu + \sigma\Phi^{-1}(\alpha) \end{aligned}$$

Reasonable values of α are small (and for sure < 0.5), so $\Phi^{-1}(\alpha) < 0$.

Additional results on other investment rules

These calculations were derived in Kan and Zhou (2006).

- Two-fund rule of Kan and Zhou:

$$\theta_{KZ2} = \frac{1}{\eta}\left[\left(\frac{(T-N-1)(T-N-4)}{T(T-2)}\right)\left(\frac{\gamma^2}{\gamma^2 + N/T}\right)\right]\hat{\Sigma}^{-1}\hat{\mu}$$

with $\gamma^2 = \tilde{\mu}'\Sigma^{-1}\tilde{\mu}$. Kan and Zhou (2006) recommend the following feasible rule $\hat{\theta}_{KZ2}$ where γ^2 is replaced by

$$\hat{\gamma}_a^2 = \frac{(T-N-2)\hat{\gamma}^2 - N}{T} + \frac{2(\hat{\gamma}^2)^{N/2}(1+\hat{\gamma}^2)^{-(T-2)/2}}{TB_{\hat{\gamma}^2/(1+\hat{\gamma}^2)}(N/2,(T-N)/2)}$$

with $\hat{\gamma}^2 = \hat{\mu}'\hat{\Sigma}^{-1}\hat{\mu}$ and $B_x(a,b)$ is the incomplete beta function

$$B_x(a,b) = \int_0^x y^{a-1}(1-y)^{b-1}dy$$

- Three-fund rule of Kan and Zhou:

$$\theta_{KZ3} = \frac{c_3}{\eta}\left[\left(\frac{\psi^2}{\psi^2 + N/T}\right)\hat{\Sigma}^{-1}\hat{\mu} + \left(\frac{N/T}{\psi^2 + N/T}\right)\mu_g\hat{\Sigma}^{-1}\iota\right]$$

with

$$\begin{aligned}
\mu_g &= \frac{\iota'\hat{\Sigma}^{-1}\hat{\mu}}{\iota'\hat{\Sigma}^{-1}\iota} \\
c_3 &= \left(\frac{T-N-4}{T}\right)\left(\frac{T-N-1}{T-2}\right) \\
\psi^2 &= (\mu - \mu_g\iota)'\Sigma^{-1}(\mu - \mu_g\iota)
\end{aligned}$$

Kan and Zhou (2006) recommend the following feasible rule $\hat{\theta}_{KZ3}$ where μ_g and ψ^2 are respectively replaced by:

$$\begin{aligned}
\hat{\mu}_g &= \frac{\hat{\mu}'\hat{\Sigma}^{-1}\iota}{\iota'\hat{\Sigma}^{-1}\iota} \\
\hat{\psi}^2_a &= \frac{(T-N-1)\hat{\psi}^2 - (N-1)}{T} + \frac{2(\hat{\psi}^2)^{(N-1)/2}(1+\hat{\psi}^2)^{-(T-2)/2}}{TB_{\hat{\psi}^2/(1+\hat{\psi}^2)}((N-1)/2,(T-N+1)/2)}
\end{aligned}$$

- Sequential min-max of Garlappi, Uppal and Wang:

$$\theta_{GUW} = \frac{1}{\eta}d\frac{T-1}{T}\hat{\Sigma}^{-1}\hat{\mu}$$

$$\text{where} \quad d = \begin{cases} 1 - (\epsilon/\hat{\gamma}^2)^{1/2} & if \quad \hat{\gamma}^2 > \epsilon \\ 0 & if \quad \hat{\gamma}^2 \leq \epsilon \end{cases} \qquad \text{with} \quad \epsilon = N\mathcal{F}^{-1}_{N,T-N}(p)/(T-N)$$

where $\mathcal{F}^{-1}_{N,T-N}$ is the inverse of the cumulative distribution function of a central F-distribution with $(N, T-N)$ degrees of freedom and p is a probability. We use $p = .99$ as suggested in Garlappi *et al.* (2006).

- Expected performances of the investment rules considered in section 5. See also Kan and Zhou (2006) for additional details.

(1) Parameter certainty optimal: $E_1 = \gamma^2/(2\eta)$.

(1f) Feasible counterpart of (1):

$$E_{1f} = k_1 \frac{\gamma^2}{2\eta} - \frac{NT(T-2)}{2\eta(T-N-1)(T-N-2)(T-N-4)}$$

$$\text{with } k_1 = \left(\frac{T}{T-N-2}\right)\left[2 - \frac{T(T-2)}{(T-N-1)(T-N-4)}\right]$$

(2) Optimal 2-fund rule:

$$E_2 = \frac{\gamma^2}{2\eta} \times \frac{(T-N-4)(T-N-1)}{(T-2)(T-N-2)} \times \frac{\gamma^2}{\gamma^2 + N/T}$$

(2f) Feasible counterpart of (2): E_{2f} must be evaluated by simulation.

(3) Optimal 2-fund rule with known variance:

$$E_3 = \frac{\gamma^2}{2\eta} \times \frac{\gamma^2}{\gamma^2 + N/T} \times \frac{T}{T-N-2}\left[2 - \frac{(T+N)(T-2)}{(T-N-1)(T-N-4)}\right]$$

(3f) Feasible counterpart of (3): E_{3f} must be evaluated by simulation.

(4) Bayesian with diffuse priors:

$$E_4 = k_2 \frac{\gamma^2}{2\eta} - \frac{NT(T-2)(T-N-2)}{2\eta(T+1)^2(T-N-1)(T-N-4)}$$

$$\text{with } k_2 = \frac{T}{T+1}\left[2 - \frac{T(T-2)(T-N-2)}{(T+1)(T-N-1)(T-N-4)}\right]$$

(5) Uncertainty aversion rule: E_5 must be evaluated by simulation.

(6) Optimal 3-fund rule:

$$E_6 = \frac{\gamma^2}{2\eta}\frac{(T-N-1)(T-N-4)}{(T-2)(T-N-2)}\left[1 - \frac{N/T}{\gamma^2 + \left(\frac{\gamma^2}{\omega^2}\right)\left(\frac{N}{T}\right)}\right]$$

(6f) Feasible counterpart of (6): E_{6f} must be evaluated by simulation.

(7) P-value maximization given c: $E_7(c)$ must be evaluated by simulation.

Proofs of the main results

Proof of equation (3.5) $\theta_p(c)$:
The first order conditions can be reinterpreted as a function of the (feasible) vector of the mean-variance weights $\hat{\theta}_{MV}$ defined in equation (2.2) as follows:

$$(\hat{\bar{\mu}} - \eta\hat{\Sigma}\theta_p)\sqrt{\theta_p'\hat{\Sigma}\theta_p} - \frac{\theta_p'\hat{\bar{\mu}} - (\eta/2)(\theta_p'\hat{\Sigma}\theta_p) - c}{\sqrt{\theta_p'\hat{\Sigma}\theta_p}}\hat{\Sigma}\theta_p = 0$$

$$\Leftrightarrow \quad \hat{\bar{\mu}} - \eta\hat{\Sigma}\theta_p - \frac{\theta_p'\hat{\bar{\mu}} - c}{\theta_p'\hat{\Sigma}\theta_p}\hat{\Sigma}\theta_p + \frac{\eta}{2}\hat{\Sigma}\theta_p = 0$$

$$\Leftrightarrow \quad \hat{\bar{\mu}} - \frac{\eta}{2}\hat{\Sigma}\theta_p - \frac{\theta_p'\hat{\bar{\mu}} - c}{\theta_p'\hat{\Sigma}\theta_p}\hat{\Sigma}\theta_p = 0$$

$$\Leftrightarrow \quad \hat{\Sigma}^{-1}\hat{\bar{\mu}} - \frac{\eta}{2}\theta_p - \frac{\theta_p'\hat{\bar{\mu}} - c}{\theta_p'\hat{\Sigma}\theta_p}\theta_p = 0$$

$$\Leftrightarrow \quad \eta \times \hat{\theta}_{MV} = \left[\frac{\theta_p'\hat{\bar{\mu}} - c}{\theta_p'\hat{\Sigma}\theta_p} + \frac{\eta}{2}\right]\theta \tag{A.1}$$

Now for a given threshold c, we can always define a constant real number k_c such that:

$$k_c \times \eta = \frac{\theta_p'\hat{\bar{\mu}} - c}{\theta_p'\hat{\Sigma}\theta_p} \tag{A.2}$$

Then, substituting (A.2) into (A.1) yields to:

$$\eta \times \hat{\theta}_{MV} = \left(k_c + \frac{1}{2}\right) \times \eta \times \theta_p \quad \Leftrightarrow \quad \theta_p = \frac{1}{k_c + 1/2}\frac{1}{\eta}\hat{\Sigma}^{-1}\hat{\bar{\mu}} \tag{A.3}$$

If I consider $\hat{\theta}_{MV}$ as a function of η such that $\hat{\theta}_{MV} = (\hat{\Sigma}^{-1}\tilde{R})/\eta$ and θ_p as a function of η (where θ_p is the weighting vector maximizing the p-value of the test with a parameter η of risk-aversion), then I get:

$$\theta_p(\eta) = \hat{\theta}_{MV}(\tilde{\eta}) \quad \text{with} \quad \tilde{\eta} = \eta \times (k_c + \frac{1}{2}) \tag{A.4}$$

The interpretation of $\tilde{\eta}$ as a *corrected* parameter of risk aversion is valid if and only if $k_c + 1/2 > 0$. This result may appear a bit *ad hoc* at first because k_c depends on the unknown

vector of weights θ. But from equation (A.2), we are actually able to deduce its explicit expression as a function of known quantities only:

$$(A.2) \iff \theta_p'\hat{\tilde{\mu}} - c = \theta_p'\tilde{\Sigma}\theta_p \times k_c \times \eta$$

Then after replacing θ_p by its expression (A.3), we get:

$$\frac{1}{k_c+1/2}\frac{1}{\eta}\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\hat{\tilde{\mu}} - c = \frac{k_c}{(k_c+1/2)^2}\frac{1}{\eta}\hat{\tilde{\mu}}\tilde{\Sigma}^{-1}\hat{\tilde{\mu}} \quad \Rightarrow \quad c = \frac{\hat{\tilde{\mu}}'\hat{\Sigma}^{-1}\hat{\tilde{\mu}}}{2\eta(k_c+1/2)^2}$$

$$\Rightarrow \quad (k_c+1/2) = \sqrt{\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}\hat{\tilde{\mu}}}{2\eta}} \times \frac{1}{\sqrt{c}}$$

Proof of equation (3.6) c^*:

$$Q(\theta_p(c)) \;=\; \theta_p'(c)\tilde{\mu}_0 - \frac{\eta}{2}\theta_p'(c)\Sigma_0\theta_p(c) = \frac{\sqrt{2c}}{\sqrt{\eta\tilde{\gamma}^2}}\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\tilde{\mu}_0 - \frac{c}{\tilde{\gamma}^2}\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\Sigma_0\tilde{\Sigma}^{-1}\hat{\tilde{\mu}}$$

where $\tilde{\gamma}^2 = \hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\hat{\tilde{\mu}}$. We then maximize it with respect to c:

$$\max_{c\geq 0} E\left[Q_P(\theta_p(c))\right]$$

The associated first order conditions are:

$$\frac{1}{\sqrt{2c^*\eta}}E\left(\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\tilde{\mu}_0}{\sqrt{\tilde{\gamma}^2}}\right) = E\left(\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\Sigma_0\tilde{\Sigma}^{-1}\hat{\tilde{\mu}}}{\tilde{\gamma}^2}\right) \Longrightarrow c^* = \frac{1}{2\eta} \times \frac{\left[E\left(\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\tilde{\mu}_0}{\sqrt{\tilde{\gamma}^2}}\right)\right]^2}{\left[E\left(\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\Sigma_0\tilde{\Sigma}^{-1}\hat{\tilde{\mu}}}{\tilde{\gamma}^2}\right)\right]^2}$$

The associated optimal vector of weights is the following:

$$\theta_p(c^*) = \frac{\left|E\left(\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\tilde{\mu}_0}{\sqrt{\tilde{\gamma}^2}}\right)\right|}{\left|E\left(\frac{\hat{\tilde{\mu}}'\tilde{\Sigma}^{-1}\Sigma_0\tilde{\Sigma}^{-1}\hat{\tilde{\mu}}}{\tilde{\gamma}^2}\right)\right|}\frac{1}{\sqrt{\tilde{\gamma}^2}}\frac{1}{\eta}\tilde{\Sigma}^{-1}\hat{\tilde{\mu}}$$

Note in particular that if $\tilde{\mu}_0$ and Σ_0 were known, we would get $\theta_p(c^*) = \theta_{MV}$, which corresponds to the best portfolio rule in absence of estimation risk.

	Mean	Standard deviation
France	0.014	0.069
Germany	0.013	0.059
Japan	0.011	0.067
UK	0.015	0.073
USA	0.012	0.044

and $\rho_0 = \begin{pmatrix} 1 & .590 & .390 & .541 & .456 \\ & 1 & .338 & .424 & .347 \\ & & 1 & .342 & .221 \\ & & & 1 & .506 \\ & & & & 1 \end{pmatrix}$

Table IV.1: Summary statistics and matrix of correlations for the MSCI of G5 countries over the period January 1974 to December 1998.

	Size of the rolling window T				
	60	120	180	240	300
Optimal benchmark c^*	0.0765	0.1415	0.1840	0.2124	0.2325

Table IV.2: Optimal benchmark c^* (in percentages per month) for several sample sizes of the rolling window used to calculate the estimators of the first two moments of the portfolio distribution. c^* has been evaluated by simulation with $M = 50000$ replications.

Rule		Size of the rolling window T				
		60	120	180	240	300
(1)	θ_{MV}	0.3503	0.3503	0.3503	0.3503	0.3503
(1f)	$\hat{\theta}_{MV}$	-0.8977	-0.1667	0.0256	0.1146	0.1648
(2)	θ_{KZ2}	0.0929	0.1518	0.1888	0.2141	0.2326
(2f)*	$\hat{\theta}_{KZ2}$	-0.0046	0.1033	0.1510	0.1832	0.2067
(3)	θ_{HRW}	0.0741	0.1418	0.1816	0.2084	0.2279
(3f)*	$\hat{\theta}_{HRW}$	-0.2577	0.0518	0.1371	0.1813	0.2090
(4)	θ_B	0.0028	0.0014	0.0010	0.0007	0.0006
(5)*	θ_{GLW}	0.0036	0.0121	0.0223	0.0356	0.0511
(6)	θ_{KZ3}	0.2827	0.3007	0.3074	0.3113	0.3140
(6f)*	$\hat{\theta}_{KZ3}$	0.0266	0.1770	0.2274	0.2530	0.2683
(7)*	$\theta_p(c_1)$	0.0835	0.1167	0.1333	0.1439	0.1509
(7)*	$\theta_p(c_2)$	0.0690	0.1545	0.1949	0.2204	0.2374
(7)*	$\theta_p(c_3)$	-0.0050	0.1190	0.1761	0.2117	0.2352
(7)*	$\theta_p(c^*)$	0.0933	0.1564	0.1950	0.2223	0.2417

Table IV.3: Expected performances (in percentages per month) for several sample sizes of the rolling window used to calculate the required estimators of the first two moments of the portfolio distribution. A star (*) identifies a rule whose expected performance has been evaluated through a simulation with $M = 50000$ replications. For $\hat{\theta}_{KZ2}$ and $\hat{\theta}_{KZ3}$ we follow the recommendations of Kan and Zhou (2006); for θ_{GLW} we follow Garlappi *et al.* (2006). The benchmarks are chosen as $c_1 = 0.3503$, $c_2 = .1751$ and $c_3 = .3153$; for the optimal c^*, see table IV.2.

T	60		120		180		240		300	
$\hat{\theta}_{KZ2}$	105.0	(101.3)	32.0	(70.5)	20.0	(56.9)	14.4	(47.7)	11.1	(41.0)
$\hat{\theta}_{HRW}$	447.8	(173.6)	63.5	(85.2)	24.5	(60.9)	13.0	(48.2)	8.3	(40.3)
$\hat{\theta}_{KZ3}$	90.6	(92.4)	41.1	(49.5)	26.0	(35.1)	18.7	(27.8)	14.5	(23.4)

Table IV.4: Expected performance losses (in percentages per month) when using the feasible rule instead of its theoretical counterpart. For convenience, we also report in parentheses the loss of using the feasible rule instead of the true optimal one θ_{MV}.

Rule		Size of the rolling window T				
		60	120	180	240	300
(1)	θ_{MV}	0	0	0	0	0
(1f)	$\hat{\theta}_{MV}$	27.9694	12.6632	8.1790	6.0430	4.7878
(2)	θ_{KZ2}	6.5508	5.1686	4.2363	3.5860	3.1047
(2f)	$\hat{\theta}_{KZ2}$	11.3447	6.5516	4.8417	3.9102	3.3040
(3)	θ_{HRW}	8.2777	5.7838	4.5618	3.7894	3.2443
(3f)	$\hat{\theta}_{HRW}$	18.5680	8.8222	5.9800	4.5882	3.7469
(4)	θ_B	24.3013	11.8256	7.8175	5.8424	4.6606
(5)	θ_{GHW}	0.5518	0.5340	0.6264	0.6942	0.7478
(6)	θ_{KZ3}	3.8310	2.0374	1.4739	1.2097	1.0583
(6f)	$\hat{\theta}_{KZ3}$	12.7672	6.4304	4.3444	3.3113	2.6818
(7)	$\theta_p(c_1)$	4.6841	2.7357	1.9583	1.5330	1.2595
(7)	$\theta_p(c_2)$	10.4739	6.1172	4.3789	3.4278	2.8162
(7)	$\theta_p(c_3)$	14.0521	8.2071	5.8749	4.5989	3.7784
(7)	$\theta_p(c^*)$	6.9220	5.4983	4.4881	3.7747	3.2447

Table IV.5: Average transaction costs over an investment horizon $T_h = 60$. We consider several rolling window sizes (to evaluate the estimators of the first two moment of the distributions of the returns) and $M = 50000$ replications.

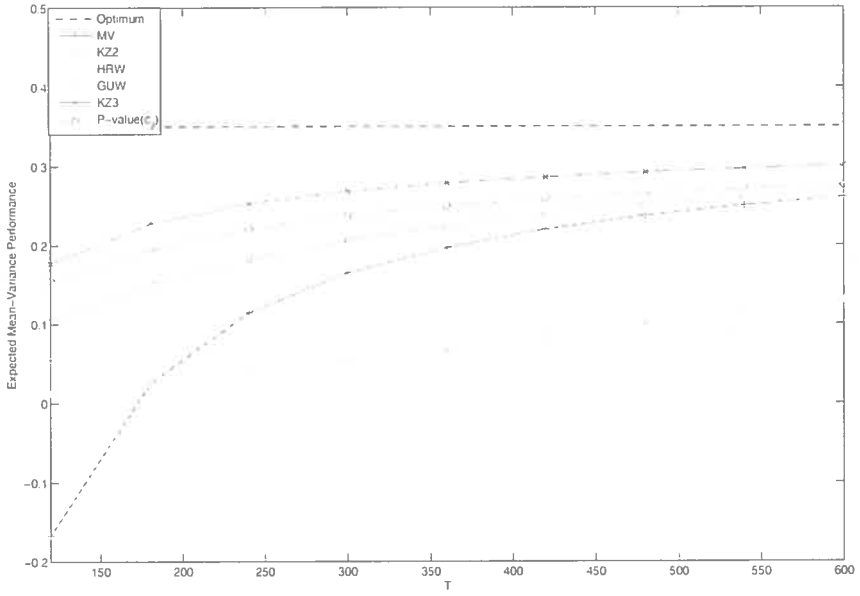

Figure IV.1: Expected performances (in percentages per month) for several feasible investment rules as a function of the size of the rolling window used to calculate the required estimators.

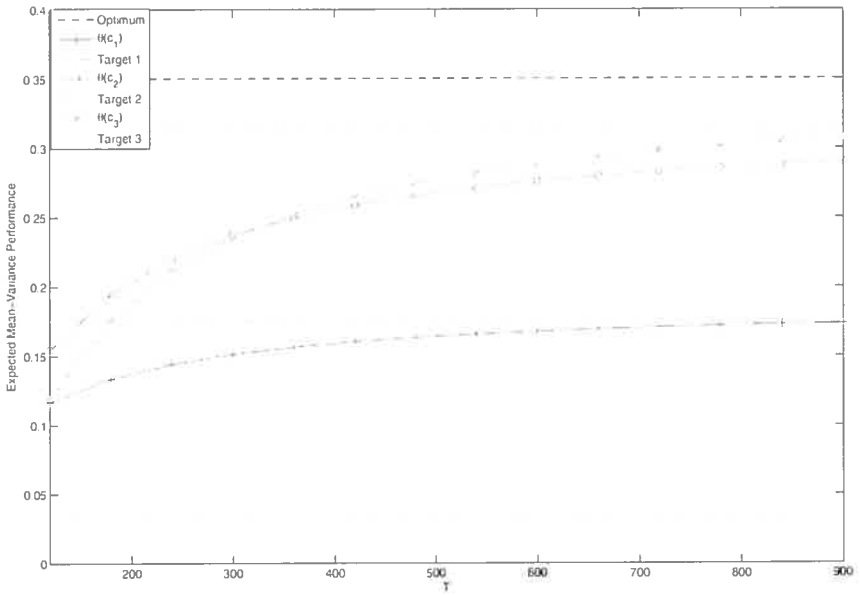

Figure IV.2: Expected performances (in percentages per month) for the p-value investment rules with several benchmarks as a function of the size of the rolling window used to calculate the required estimators.

Bibliographie

[1] K.M. Abadir and J.R. Magnus, *Matrix Algebra*, Cambridge University Press, Cambridge, 2005.

[2] S.I. Amari, *Differential-Geometrical Methods in Statistics*, 2nd ed., Lecture Notes in Statistics No 28, Berlin, 1990.

[3] D.W.K. Andrews, *Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity*, Econometrica **62** (1994), 43–72.

[4] ———, *Nonparametric Kernel Estimation for Semiparametric Econometric Models*, Econometric Theory **11** (1995), 560–596.

[5] D.W.K. Andrews and J.H. Stock, *Inference with Weak Instruments*, Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society (W.K. Newey R. Blundell and T. Persson, eds.), vol. III, Cambridge University Press, Cambridge, UK, 2007.

[6] J.D. Angrist and A. Krueger, *How compulsory School Attendance affect Schooling and Earnings*, Quarterly Journal of Economics **106** (1991), 979–1014.

[7] B. Antoine, H. Bonnal, and E. Renault, *On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood*, Journal of Econometrics **138** (2007), 461–487.

[8] B. Antoine and E. Renault, *Efficient Minimum Distance Estimation with Multiple Rates of Convergence*, Working Paper (2007).

[9] V. Bawa, S. Brown, and R. Klein, *Estimation Risk and Optimal Portfolio Choice*, North Holland, Amsterdam, 1979.

[10] M. Brandt, *Portfolio Choice Problems*, in Handbook of Financial Econometrics (2004), Y. Aït-Sahalia and L.P. Hansen, eds.

[11] M. Caner, *Testing, Estimation and Higher Order Expansions in GMM with Nearly-Weak Instruments*, Working Paper (2005).

[12] F. Cavadini, A. Sbuelz, and F. Trojani, *A simplified way of incorporating Model Risk, Estimation Risk and Robustness in Mean Variance Portfolio Management*, Working paper (2001).

[13] D. Cho, *Uncertainty in Second Moments: Implications for Portfolio Allocation*, Working paper (2003), SUNY at Buffalo.

[14] F. Critchley, P.K. Marriott, and M. Salmon, *On the Differential Geometry of the Wald Test with Nonlinear Restrictions*, Econometrica **64** (1996), 1213–1222.

[15] J. Davidson, *Econometric Theory*, Blackwell Publisher, Oxford, 2000.

[16] R. Davidson and J.G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford, 1993.

[17] S.G. Donald and W.K. Newey, *A Jackknife interpretation of the Continuous Updating Estimator*, Economics Letter **67** (2000), 239–243.

[18] P. Dovonon and E. Renault, *GMM Overidentification Test with First-Order Unidentification*, Working Paper (2006).

[19] J.M. Dufour, *Some Impossibility Theorems in Econometrics with Application to Structural and Dynamic Models*, Econometrica **65** (1997), 1365–1387.

[20] ———, *Identification, Weak instruments and Statistical Inference in Econometrics*, Canadian Journal of Economics **36** (2003), no. 4, 767–808, Presidential Address to the Canadian Economics Association.

[21] P. Gagliardini, C. Gouriéroux, and E. Renault, *Efficient Derivative Pricing by Extended Method of Moments*, Working Paper (2005).

[22] L. Garlappi, R. Uppal, and T. Wang, *Portfolio selection with Parameter and Model Uncertainty: A Multi-Prior Approach*, The Review of Financial Studies **20** (2007), 41–81.

[23] C. Gouriéroux and A. Monfort, *Statistics and Econometric Models*, Cambridge University Press, Cambridge, 1995.

[24] A.W. Gregory and M.R. Veall, *On Performing Wald Tests for Nonlinear Restrictions*, Econometrica **53** (1985), 1465–1468.

[25] J. Hahn and G. Kuersteiner, *Discontinuities of Weak Instruments limiting Distributions*, Economics Letters **75** (2002), 325–331.

[26] A.R. Hall, *Generalized Method of Moments*, Advanced Texts in Econometrics, Oxford University Press, 2005.

[27] L.P. Hansen, *Large Sample Properties of Generalized Method of Moments Estimators*, Econometrica **50** (1982), no. 4, 1029–1054.

[28] L.P. Hansen, J. Heaton, and A. Yaron, *Finite Sample Properties of some Alternative GMM Estimators*, Journal of Business and Economic Statistics **14** (1996), 262–280.

[29] C.R. Harvey, J. Liechty, M. Liechty, and P. Müller, *Portfolio Selection with Higher Moments*, Working paper (2004), Duke university.

[30] J. Horst, F. de Roon, and B. Werker, *An Alternative Approach to Estimation Risk*, in Advances in Corporate Finance and Asset Pricing, Elsevier Amsterdam (2006), Luc Renneboog, eds.

[31] P. Jorion, *Value-at-Risk: the new Benchmark for controlling Market Risk*, second ed., McGraw-Hill, Chicago, Irwin, 1996.

[32] R. Kan and G. Zhou, *Optimal Portfolio Choice with Parameter Uncertainty*, Journal of Financial and Quantitative Analysis (2007), forthcoming.

[33] F. Kleibergen, *Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression*, Econometrica **70** (2002), 1781–1803.

[34] ———, *Testing Parameters in GMM without assuming that they are identified*, Econometrica **73** (2005), 1103–1123.

[35] N. Kocherlakota, *On Tests of Representative Consumer Asset Pricing Models*, Journal of Monetary Economics **26** (1990), 285–304.

[36] L. Lee, *Pooling Estimates with Different Rates of Convergence - A minimum χ^2 Approach: with an emphasis on a Social Interactions Model*, Working Paper (2004), Ohio State University.

[37] ———, *Classical Inference with ML and GMM Estimates with Various Rates of Convergence*, Working Paper (2005), Ohio State University.

[38] H.M. Markowitz, *Portfolio Selection: Efficient Diversification of Investments*, Wiley, New-York, 1959.

[39] P. Marriott and M. Salmon, *Applications of Differential Geometry to Econometrics*, Cambridge University Press, 2000.

[40] A. Meucci, *Risk and Asset Allocation*, Springer-Verlag, New-York, 2005.

[41] M. Moreira, *A Conditional Likelihood Ratio Test for Structural Models*, Econometrica **71** (2003), no. 4, 1027–1048.

[42] W.K. Newey, *Generalized Method of Moments Specification Testing*, Journal of Econometrics **29** (1985), 229–256.

[43] W.K. Newey and R.J. Smith, *Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators*, Econometrica **72** (2004), 219–255.

[44] W.K. Newey and K.D. West, *Hypothesis Testing with Efficient Method of Moments Estimation*, International Economic Review **28** (1987), 777–787.

[45] P.C.B. Phillips, *Partially Identified Econometric Models*, Econometric Theory **5** (1989), 181–240.

[46] P.C.B. Phillips and J.Y. Park, *On the Formulation of Wald Tests of Nonlinear Restrictions*, Econometrica **56** (1988), 1065–1083.

[47] J.D. Sargan, *Identification and Lack of Identification*, Econometrica **51** (1983), 1605–1634.

[48] D. Staiger and J.H. Stock, *Instrumental Variables Regression with Weak Instruments*, Econometrica **65** (1997), no. 3, 557–586.

[49] J.H. Stock and J.H. Wright, *GMM with Weak Identification*, Econometrica **68** (2000), no. 5, 1055–1096.

[50] J.H. Stock, J.H Wright, and M. Yogo, *A Survey of Weak Instruments and Weak Identification in Generalized Methods of Moments*, Journal of Business and Economic Statistics **20** (2002), 518–529.

[51] M. Stutzer, *Portfolio Choice with Endogenous Utility: a Large Deviations Approach*, Journal of Econometrics **116** (2003), 365–386.

[52] G. Tauchen, *Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data*, Journal of Business and Economic Statistics **4** (1986), 397–425.

[53] G. Tauchen and R. Hussey, *Quadrature-based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models*, Econometrica **59** (1991), 371–396.

[54] A.W. van der Vaart, *Asymptotic Statistics*, Series in Statistical and Probabilistic Mathematics ed., Cambridge University Press, Cambridge, 1998.

[55] A. Wald, *Tests of Statistical Hypotheses concerning Several Parameters when the Number of Observations is Large*, Transactions of the American Mathematical Society **54** (1943), 426–482.

[56] J. Wang and E. Zivot, *Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments*, Econometrica **66** (1998), no. 6, 1389–1404.

[57] A. Zellner and V.K. Chetty, *Prediction and Decision Problems in Regression Models from the Bayesian Point of View*, Journal of the American Statistical Association **60** (1065), 608–616.

Conclusion générale

Dans cette thèse, nous avons étudié des contextes d'application, en particulier dans le domaine de l'économie financière, où le point de vue asymptotique traditionnel pouvait être trompeur. Chacun des quatre essai a alors proposé une méthode pour affiner les approximations asymptotiques en présence d'échantillons d'observations, toujours finis en pratique.

Dans le premier essai, nous avons proposé un cadre de travail général, dans un contexte GMM, afin de tenir compte d'instruments potentiellement faibles. En contraste avec la littérature existante, la faiblesse a directement été mise en relation avec les conditions de moment (à travers les instruments) et plus seulement avec les paramètres. Plus précisément, nous avons considéré deux groupes de conditions de moment : le groupe standard, associé au taux de convergence usuel $\sqrt{T}$ et le groupe faible, associé à un taux de convergence plus lent λ_T. Ce cadre garantit la convergence des estimateurs GMM de tous les paramètres, mais à des taux possiblement plus lents que d'habitude. De plus, nous avons identifié et estimé des directions dans l'espace des paramètres, qui convergent à la vitesse standard $\sqrt{T}$. Par ailleurs, nous avons également caractérisé la validité des approches de test standard, comme les tests de Wald et GMM-LM. De tels résultats sont d'un intérêt pratique certain, puisque la connaissance du taux de convergence lent n'est pas requise.

La simulation d'un modèle d'équilibre général CCAPM a révélé que les estimateurs GMM convergeaient tous. De plus, dans certaines configurations des valeurs des paramètres, leurs taux de convergence apparaissaient plus lents que le traditionnel $\sqrt{T}$, tandis qu'une combinaison linéaire des paramètres structurels était estimée au taux standard. Dans les modèles plus sophistiqués que le CCAPM (par exemple, les modèles avec préférence Epstein-Zin, formation d'habitude ou encore évaluation d'options), la distinction entre directions fortement

identifiées dans l'espace des paramètres et directions faiblement identifiées peut également compléter notre interprétation économique du modèle.

Afin de simplifier les notations et l'exposition, nous avons choisi de nous concentrer ici sur deux groupes de conditions de moment, seulement. L'extension vers de multiples groupes, et donc différents degrés de faiblesse, est naturelle : pensons, par exemple, à un praticien qui utiliserait des instruments de différentes qualités informationnelles. Un autre document de travail (Antoine et Renault (2007)), en préparation, considère ce problème plus spécifiquement. Toutefois, aucune application reliée à des problèmes d'identification n'y est envisagée ; ce point de vue est spécifique à cet essai.

Dans le deuxième essai, nous avons réalisé une étude comparative de puissance entre le test standard GMM-LM et sa correction proposée par Kleibergen (2005). Nous avons montré que cette correction avait des conséquences asymptotiques en présence de problèmes d'identification : en particulier lorsque des instruments de qualité hétérogène sont utilisés. Nous recommandons donc l'utilisation du cadre de travail développé dans le chapitre 1. Il n'a pas beaucoup de conséquences en termes de spécification des problèmes d'identification. De plus, non seulement, il donne accès aux procédures de test standard, mais il permet aussi d'identifier les directions dans l'espace des paramètres contre lesquelles les tests ont de la puissance.

En ce qui concerne les tests sur des sous-vecteurs des paramètres, la supériorité du cadre de travail du chapitre 1 est claire. En particulier, la reparamétrisation permet d'identifier les directions pour lesquelles le test GMM-LM standard s'applique directement. Aucune hypothèse supplémentaire sur les paramètres non-testés n'est nécessaire. Pour finir, des transformations (non-linéaires) générales des paramètres peuvent également être testées dans ce contexte. Ceci n'est pas évoqué dans Kleibergen (2005).

Dans le troisième essai, nous avons proposé une nouvelle méthode d'inférence, la procédure Modified-Wald, afin de fournir de l'inférence fiable sur un ratio de paramètres multidimensionnel. Cette nouvelle méthode est basée sur la statistique de Wald et démontre la même commodité computationnelle. En plus, elle est associée à des régions de confiance non bornées lorsque l'identification fait défaut, comme suggéré par Dufour (1997). Notre

idée consiste à intégrer le contenu informationnel de l'hypothèse nulle dans le calcul de sa métrique.

Nous avons montré que la statistique Modified-Wald est asymptotiquement équivalente à la statistique de Wald. Sa région de confiance, au niveau α, est non bornée avec une probabilité aussi élevée que (1-α), la borne supérieure déduite dans Dufour (1997). Ces résultats ont été appliqués à deux cas d'étude. L'exercice de simulation associé à un ratio bidimensionnel a révélé que la procédure Modified-Wald était capable de détecter une situation dans laquelle l'information échantillonnale n'était pas suffisante pour fournir de l'inférence précise. La seconde application, sur le modèle de régression linéaire avec variables instrumentales, nous a permis de construire un pont entre la perte d'identification à la frontière de l'espace des paramètres et sa modélisation économétrique (artificielle) à travers la taille de l'échantillon.

Dans le quatrième essai, nous avons proposé une nouvelle façon de tenir compte du risque d'estimation en sélectionnant le portefeuille optimal. En contraste avec la littérature existante, le portefeuille optimal est défini de manière plus conservative, en cherchant (seulement) à garantir une performance minimale, et plus à maximiser la performance directement. Plus précisément, notre méthode de sélection repose sur un test unilatéral qui assure que la performance du portefeuille est au-dessus d'un seuil donné. Les poids optimaux respectifs des actifs financiers sont alors obtenus à partir de la maximisation de la p-valeur associée à ce test. Ce test nous a permis de définir une méthode qui intègre directement le risque d'estimation. De plus, en négligeant le risque d'estimation de la variance, nous avons obtenu une règle d'investissement explicite directement applicable.

Nos simulations ont montré que, pour des choix de référence raisonnables, la performance du portefeuille associé à la règle d'investissement p-valeur était très satisfaisante, surtout avec de petits échantillons ; de plus, les coûts de maintenance associés étaient généralement faibles, ce qui témoigne de la stabilité de la règle à travers le temps.

Pour finir, plusieurs voies de recherche peuvent être envisagées. On peut tout d'abord penser à introduire des performances de référence aléatoires : cela permettrait de traquer directement la performance de certains indices financiers d'intérêt. On peut aussi chercher à introduire le risque de modèle dans le choix du portefeuille optimal, à l'image de Cavadini, Sbuelz et Trojani (2001).