Essays in Theoretical and Applied Econometrics

Di Liu

A Thesis

In the Department

of

Economics

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Economics) at

Concordia University

Montreal, Quebec, Canada

June 2014

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Di Liu

Entitled: Essays in Theoretical and Applied Econometrics

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Economics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Michael Cornway — Chair

Valentyn Panchenko — External Examiner

Yogen Chaubey — External to Program

Gordon Fisher — Examiner

Bryan Campbell — Examiner

Artem Prokhorov, Prosper Dovonon — Thesis Supervisor

Approved by

Chair of Department or Graduate Program Director

Dean of Faculty

ABSTRACT

Essays in Theoretical and Applied Econometrics

Di Liu, Ph.D.
Concordia University, 2014

This thesis investigates three topics in theoretical and applied econometrics: two sample nonparametric estimation of intergenerational income mobility, sparse sieve maximum likelihood estimation, and asymptotic efficiency of Improved QMLE and Sieve MLE.

The first essay proposes a two sample nonparametric GMM estimator, which extends the local linear GMM estimator to two sample settings, and applies it to estimate the intergnerational income mobility in the U.S and Sweden. The second essay proposes an estimator that uses the Dantzig Selector to improve the finite sample performance of Sieve MLE in a panel data setting. We show that in simulations the sparsity imposed by the Dantzig Selector is innocuous with respect to the sieve MLE, and substantially improves its computational efficiency. The third essay compares an optimal GMM estimator, known as Improved QMLE, with sieve MLE in a panel data setting. We derive a condition when these two estimators are equally efficient asymptotically and provide simulation results to illustrate the extent of efficiency loss.

ACKNOWLEDGEMENTS

First of all, I want to thank Professor Artem Prokhorov for his exceptional guidance to my research work, his patience and approachability to answer my numourous questions, his encouragements to make me pursue my dream career when I was uncertain, his generous financial support during my Ph.D studies and many other research-related things. Professor Prokhorov not only teaches me everything about doing research, but also intrigues my own research interests which make me enjoy doing econometrics in the future. I am grateful to be able to have such a wonderful mentor.

I wish to thank Professor Irina Murtazashvili, my coauther, for her important contributions and comments in the first essay. Professor Murtazashvili not only makes a thorough analysis in the empirical part, but also gives me many constructive suggestions on the computational part.

I would also like to thank Professor Prosper Dovonon and Professor Nikolay Gospodinov for their great help and valuable advice to my research and career. Professor Dovonon has spent much time to discuss my research and share his working experience with me. Professor Gospodinov taught me several courses in advanced econometrics which always are my first reference, and gave me a great financial support during my Ph.D studies. I also wish to express my thanks to Professor Gordon Fisher and Professor Bryan Campbell for their constructive comments and helpful proof-reading of my thesis.

I also want to thank Elise, Lucy and Lise in Economics Deparments for their helps and understanding during my stay at Concordia University.

Finally, I wish to express my gratitude towards our parents for their devoted love, helpful support and great encouragement. I want to thank my wife, Mingzhu, for her great dedication to family, for her considerate understanding when I am pursuing my career, for her tender love which make me laugh everyday. Without her, I would not be able to finish my Ph.D studies.

CONTRIBUTION OF AUTHORS

Chapter 1

This paper is co-authored with Dr. Irina Murtazashili and Dr. Artem Prokhorov. The idea of this paper immerged when Dr. Murtazashili gave a talk in Montreal. In discussions following seminar, Dr. Prokhorov and Dr. Murtazashili started to develop the idea of a two-sample nonparametric estimator (TS-NPGMM). They proved the consistency and asymptotic normality of this estimator. Later on, I joined this project and worked on the computational tasks. First, I conducted Monte Carlo simulations to study the finite sample behavior of TS-NPGMM. Second, I applied the TS-NPGMM to the estimation of intergenerational income mobility in the U.S and Sweden. Third, I proposed a matching algorithm for bandwidth selection and variance estimation.

Chapter 2

This paper is co-authored with Dr. Artem Prokhorov. The idea of using the Dantzig selector to reduce the parameter dimension was first proposed by Dr. Prokhorov, when he was on research leave at Harvard University. Dr. Prokhorov proved the oracle inequality. I designed an algorithm applying the Dantzig selector to sieve maximum likelihood estimation (DS-SMLE). I compared the properties of DS-SMLE with brute-force SMLE through simulations. Finally, I illustrated the use of DS-SMLE with an insurance application.

Contents

List of Figures

List of Tables

Introduction

This thesis investigates three topics in theoretical and applied econometrics: two sample nonparametric GMM estimation of intergenerational income mobility, sparse sieve maximum likelihood estimation, and asymptotic efficiency of Improved QMLE and sieve MLE. They are organized as Chapter 1, 2 and 3 respectively.

Chapter 1 proposes an estimator which extends the local linear GMM in Cai and Li (2008) to a two sample setting, and estimates the intergenerational income mobility in the United States and Sweden using this new approach. Unlike existing measure of the degree to which earnings are transmitted from one generation to another, our estimator is nonparametric and applies when other estimators are infeasible. We allow intergenerational income mobility to depend flexibly on observable family background, which is particularly relevant for cross-country comparisons. We further allow for data on fathers and sons to come from different samples, which solves a critical missing data issue and alleviates attrition concerns. Finally, our estimator is consistent in the presence of measurement errors in father's long-run economic status. Using the US and Swedish data, we argue that previous parametric estimates of income mobility tend to conceal the heterogeneous nature of the transmission mechanism by keeping mobility constant across families. The striking differences we find between mobility patterns across family backgrounds as captured by father's education lead us to question the conventional result that intergenerational transmission of earnings is weaker in Sweden than in the United States, for important parts of population.

The Dantzig Selector (Candes and Tao (2007)) is traditionally used for point

estimation by least squares when the number of parameters exceeds the number of observations. Chapter 2 uses it to obtain smaller standard errors in a sieve maximum likelihood estimation in a panel setting. We assume correctly specified likelihood model for each cross section and the Bernstein polynomial serves as a copula sieve capturing dependence between them. This estimator has smaller standard errors asymptotically than the conventional QMLE but, in finite samples, the number of parameters in the sieve is close to the number of observations and may exceed it. At the same time, most of the sieve parameters are close to zero. We propose an estimator that uses the Dantzig Selector to find the sparsest vector of the sieve parameters satisfying the first order conditions of MLE up to a given tolerance level. We show in simulations that our estimator produces a sparse sieve MLE with finite-sample properties very similar to the non-sparse alternative, and substantially better than the QMLE. As a theoretical motivation for the good performance of sparse SMLE, we provide an oracle inequality relating the risk of the sparse estimator with that of an infeasible estimation where an oracle tells us which coefficients are insignificant. We also study the parameter path behavior for various tolerance levels and consider a version of a double Dantzig selector which resolves the arbitrariness in choosing the tolerance level.

Chapter 3 compares Improved QMLE (IQMLE) with Sieve MLE (SMLE) in a panel data setting in which we assume to have correctly specified probabilistic models for each cross sections and it is all the information we have. IQMLE, proposed by Prokhorov and Schmidt (2009a), is an optimally weighted GMM estimator based on the marginal scores. Sieve MLE, proposed by Panchenko and Prokhorov (2013), can be viewed as MLE except it uses a copula sieve to approximate the true copula. IQMLE is known to be asymptotically efficient among estimators using the marginal scores as moment conditions, while SMLE is known to reach a semiparametric efficiency bound for the marginal parameters. We interpret the variance of IQMLE and SMLE in an intuitive way, and derive a condition when they are equally efficient. The simulation results show that SMLE can be

relatively more efficient than IQMLE up to 70%.

Finally, I develop a Matlab package called *CopulasToolbox* to conduct different simulations in Chapter3 2 and 3. This toolbox provides a uniform platform for simulations in multivariate modeling using copulas. It contains Matlab classes for multivariate probability densities, which provide easy-to-use functions for data fitting via MLE and Sieve MLE, multivariate random number generations and 3-D graphical presentation of dependence structures. Appendix A describes the design, structure and implementation of *CopulasToolbox*.

Chapter 1

Two-Sample Nonparametric Estimation of Intergenerational Income Mobility

The extent of income mobility across generations has been a focus of economists' attention for a long while. Using empirical evidence from different countries, Solon (2002) provides a thorough survey of the literature going back at least to the 1980s, on how fathers' long-run economic status affects that of their sons. Cross-country comparisons of intergenerational income mobility have been emphasized in the literature since they permit the study of the character of inequality in a particular society and produce insights into whether cross-sectional and intergenerational inequalities are linked to each other (see, e.g., Björklund and Jäntti, 1997).

Our study focuses on measuring intergenerational income mobility in Sweden and the United States. Comparisons between the United States and a Scandinavian country have been viewed as particularly relevant when studying intergenerational income mobility across countries. Gustafsson (1994), Björklund and Jäntti (1997), and Bratsberg, Røed, Raaum, Naylor, Jäntti, Eriksson, and Österbacka (2007) are just a few studies focusing on intergenerational economic status transmission in these countries. The interest in these country pairs is due to the finding highlighted

in Gottschalk and Smeeding (1997) that the Scandinavian countries have the lowest annual income inequality in contrast to the United States where income inequality is high.

While conducting the cross-country comparison of intergenerational mobility between Sweden and the USA we address three major issues previously raised in the empirical literature on intergenerational mobility. The first issue – discussed, for example, by Corak and Heisz (1999) and Bratsberg et al. (2007) – is modelling nonlinearities in intergenerational income mobility. In particular, Bratsberg et al. (2007) emphasize that the elasticity of son's income with respect to father's income – a traditional measure of income mobility – may be inappropriate, depending on whether the functional relationship between father's and son's income is linear in logs. They argue that if the functional form is nonlinear, elasticities estimated using linear models may be misleading.

There were several attempts to address this concern. In particular, Bhattacharya and Mazumder (2011) suggest a new direct measure of intergenerational income mobility obtained by using a nonparametric model where the conditional transition probability of moving across income quantiles varies with individual-specific covariates. Murtazashvili (2012) proposes yet another measure of intergenerational mobility, which is based on a random coefficient model and which allows intergenerational income mobility to vary across the distribution of families.

This paper addresses the issue of nonlinearities by designing a new nonparametric estimation strategy that allows intergenerational income mobility to vary freely across the population of families. We build on the nonparametric GMM (NPGMM) estimation of Cai and Li (2008) and exploit the functional form flexibility permitted by the nonparametric estimation in order to introduce heterogeneity in intergenerational elasticities across the population of families with different observed characteristics.

By allowing for the heterogeneity we contribute to the long standing debate

about the "mysterious background factors" (see, e.g., Solon, 1999, p. 1766) that have explanatory power for children's future earnings along with parental income. Empirical studies have looked at community background characteristics, such as community origins and neighborhood characteristics where children have been raised (see, e.g., Corcoran, Gordon, Laren, and Solon, 1992; Datcher, 1982). Other studies have considered parent's characteristics other than earnings, for example, parent's education. For example, when studying the nonlinear patterns of earnings transmission in the Nordic countries (not including Sweden) and the USA, Bratsberg et al. (2007) looked at the link between sons' and parents' education and found cross-country differences similar to those between earnings. In our analysis we employ father's education as a family background characteristic behind the heterogeneous patterns of intergenerational mobility in the US and Sweden. The nonparametric nature of our econometric method allows us to introduce this family characteristic into the equation for intergenerational mobility in a very flexible nonlinear way.

The second empirical issues we address is a lack of sufficiently comparable data. Earnings data for fathers and sons are often unavailable from a single source or are subject to severe attrition. This imposes a critical missing data problem, which causes the common single-sample-based estimators, such as OLS and 2SLS, to become infeasible. For example, the Swedish Level of Living Survey supplies information on the dependent variable – sons' earnings – and family characteristics, such as father's education, in one wave and information on the independent variable – fathers' earnings – in a different, partially matched wave. When, earnings *are* available from the same source for two generations, the sample for one of the generations usually suffers from severe attrition. For example, the US Panel Study of Income Dynamics contains earnings information for matched cross sections of individuals, one generation apart, but, due to attrition, the sample of sons is usually much smaller.

Björklund and Jäntti (1997) address these data concerns by employing a two-

sample two-stage least squares (TS-2SLS) approach, initially suggested by Angrist and Krueger (1992). This method constructs estimates of individual population moments off the two samples and uses them in a standard parametric 2SLS procedure. This simple but effective approach has become quite popular among empirical economists in various economic fields (see, e.g., Arellano and Meghir, 1992; Lefranc and Trannoy, 2005). However, this is a fully parametric method which does not permit the degree of flexibility required to study the earnings transmission patterns.

We solve this problem by developing an extension of the NPGMM estimator along the lines of Angrist and Krueger (1992). The two-sample-based estimator we propose remains consistent and demonstrates a number of robustness properties when the data comes from two somewhat heterogeneous samples. We provide the asymptotic distribution of the new estimator and devise a new method for bandwidth selection and variance estimation, suitable to our data structure.

The third empirical issue we address is measurement error. It has been long emphasized that using OLS for estimating intergenerational income mobility under a measurement error in fathers' permanent economic status may result in biased estimates (see, e.g., Solon, 1992). If valid instruments are available, numerous empirical studies have proposed ways of using instrumental variables (IV) estimation to obtain consistent estimates (see, e.g., Björklund and Jäntti, 1997; Solon, 1992, and references therein). We follow earlier work on using instruments to handle measurement error but adapt that framework to a nonparametric two-sample setting. The nonparametric GMM nature of our method allows for a consistent estimation of intergenerational elasticities in the presence of measurement error in fathers' income.

To the best of our knowledge, this is the first estimator to handle the issues of nonlinearity, missing data and measurement error simultaneously. As a result, it provides a set of empirical observations that have not been previously recognized. While our interest in developing the new method is driven primarily by

the empirical concerns raised in the literature on intergenerational mobility, the econometrics of our estimator is fairly general and can be applied in settings other than intergenerational income mobility estimation.

Some of our empirical findings are more striking than others. In both Sweden and the US, we find highly nonlinear relationships between the intergenerational income elasticity and family background captured by father's education. This finding is in contrast with the conclusions of Bratsberg at al. (2007) who detect a nonlinear pattern in intergenerational mobility for the Nordic countries they consider but not for the US. We find that earnings transmission is stronger in the USA but only when fathers have less education than at least some years of college. Earnings transmission is stronger in Sweden and quickly grows with father's education for sons of well-educated fathers. This is contrary to the conventional result that intergenerational mobility cannot be lower in the USA than in Sweden.[1] Importantly, we report the profiles of income mobility across the entire distribution of family backgrounds in the two countries, which is key to targeted policy analysis and to cross-country comparisons.

The rest of the paper is organized as follows. Section 1.1 sets the stage by describing our data and the empirical issues it brings. Section 1.2 introduces the empirical model of intergenerational mobility. Section 1.3 describes our two-sample nonparametric GMM (TS-NPGMM) estimator, discusses its properties and provides simulation results illustrating how it can handle the limitations of the data in realistic sample sizes. Appendix 1.6 contains the econometric proofs underlying Section 1.3. Section 1.4 contains a discussion of our empirical findings and policy implications. Section 1.5 concludes.

[1] A notable exception is Jantti, Bratsberg, Roed, Raaum, Naylor, Osterbacka, Bjorklund, and Eriksson (2006) who find that mobility out of the lowest quintile of the income distribution is much lower in the US, while in the Nordic countries income persistence is higher in the upper tails of the distribution. We thank a referee for providing this reference to us.

1.1 Data Description

Following Björklund and Jäntti (1997), we use the Panel Study of Income Dynamics (PSID) and the Swedish Level of Living Survey (SLLS) to obtain data for the United States and Sweden, respectively. Both, the PSID and SLLS are longitudinal (but not necessarily annual) surveys conducted since 1968. As Björklund and Jäntti (1997) point out, while in its original wave of 1968 the SLLS was based on a representative sample of around 6,000 individuals, new individuals were added to the sample in the following waves to ensure representativeness for the whole population. Therefore, by using a two-sample approach we are able to exploit all observations available in the SLLS.

Furthermore, while the PSID is a longitudinal survey following individuals and their families from the original wave of 1968 onward, the two-sample approach using data from the PSID permits, similarly to the SLLS, to handle the missing data issues more adequately than single-sample methods. Specifically, using a two-sample approach to study the relation between fathers' and sons' incomes, we are able to circumvent the attrition problem in the PSID. Besides using observations on individuals added to the survey in the later waves, we are able to exploit observations on spouses of grown children that were a part of the survey from its origination (as long as they report information on their fathers' education), which we cannot do in a one-sample setting.

We create our US and Swedish samples using the guidelines from Björklund and Jäntti (1997). Specifically, our US sample of individuals – we will refer to these individuals as *fathers* – is taken from the 1968 wave and contains 1,613 male heads of household of age between 27 and 68 who had at least one child (daughter or son). This sample is obtained from the Survey Research Center (SRC) component of the PSID. The independent sample of the US individuals – we will refer to these individuals as *sons* – taken from the 1988 SRC contains 467 individuals. Sons are restricted to those individuals who were born between 1951 and 1959 and who were the oldest sons from multiple-son families. Only those fathers and sons who

Table 1.1: Summary Statistics for US and Swedish Samples

Variables	Mean	St Dev	Min	Max
Panel A: US sample (467 sons and 1,613 fathers)				
Father's age in 1967	45.19	10.94	27	68
Father's earnings in 1967 (in 1987 $)	28,311	19,432	442	222,744
Father's log earnings in 1967	10.04	0.74	6.09	12.31
Father's education	11.45	4.12	0	18
Son's age in 1987	32.49	2.46	28	36
Son's earnings in 1987	28,598	19,352	1,200	210,000
Son's log earnings in 1987	10.06	0.67	7.09	12.25
Reported father's education	11.77	3.50	0	18
Panel B: Swedish sample (324 sons and 565 fathers)				
Father's age in 1967	42.88	7.63	25	60
Father's earnings in 1967 (in 1990 Kronas)	16,821	8,903	5,678	76,687
Father's log earnings in 1967	9.63	0.42	8.64	11.25
Father's education	8.00	2.93	6	16
Son's age in 1990	34.44	3.07	30	39
Son's earnings in 1990	1,797	637	550	5,550
Son's log earnings in 1990	7.43	0.36	6.31	8.62
Reported father's education	8.01	3.05	6	16

reported positive annual earnings for 1967 and 1987, respectively, are included into the samples. The US fathers report their 1967 annual income, education and occupation, while the US sons report their annual income in 1987, as well as occupation and education of their actual fathers.

Our Swedish sample of *fathers* (analogously to the US sample) is taken from the 1968 wave of the SLLS and contains 565 individuals who are either native Swedes or moved to Sweden before the age of 16. The sample of Swedish *sons* (by analogy to the US sample) is taken from the 1991 wave of the SLLS and contains 324 individuals who were born between 1952 and 1961. The Swedish fathers report their 1967 annual income, occupation and the highest education level attained. The Swedish sons provide information on their 1990 annual income, as well as their actual fathers' education and occupation.

Panel A of Table 1.1 presents summary statistics for fathers and sons from the

Table 1.2: Education and Occupation Characteristics of Fathers

	Fathers' own report of fathers' characteristics	Sons'
Panel A: US Sample		
Fraction with education higher than compulsory	0.94	0.97
Pearson χ^2 test p-values	0.000 to 0.078	
Fraction with given occupation:		
1 Professional, technical and kindred workers	0.16	0.15
2 Managers, officials and proprietors	0.11	0.08
3 Self-employed businessmen	0.07	0.03
4 Clerical and sales workers	0.11	0.11
5 Craftsmen, foremen, and kindred workers	0.23	0.25
6 Operatives and kindred workers	0.17	0.19
7 Laborers and service workers, farm laborers	0.09	0.08
8 Farmers and farm managers	0.04	0.09
Panel B: Swedish Sample		
Fraction with education higher than compulsory	0.62	0.63
Pearson χ^2 test p-value	0.8102	
Fraction with given occupation:		
1 Higher-grade professional	0.09	0.11
2 Lower-grade professional	0.12	0.08
3 Non-manual workers and lower-grade technicians	0.14	0.15
4 Small proprietors with employees	0.06	0.07
5 Small proprietors without employees	0.04	0.05
6 Farmers, self-employed in primary agricultural production and other workers	0.11	0.13
7 Skilled manual workers	0.22	0.18
8 Semi-skilled manual workers	0.17	0.20

US samples, while Panel B of Table 1.1 reports those for the Swedish samples. Table 1.2 presents more detailed information on education and occupation of fathers. Implicitly, there are two types of fathers in our analysis: (1) individuals observed directly in the samples of fathers (we can think of them as pseudo-fathers), and (2) individuals who were not directly observed but described by their sons in the samples of sons (we can think of them as actual fathers).

As Table 1.2 suggests, some features of the two distributions of the educational and occupational characteristics of pseudo-fathers and actual fathers appear to be reasonably close. However, a more careful consideration is warranted here.[2] When we consider the differences in all parts of the distribution, not just at a single point, the picture may change. The distribution of father's education is what matters here as this will capture the family background in what follows. Indeed the actual fractions of the observed education levels (not reported here) have more variation between the two samples than the means. When we use a two-sample Pearson Chi-square distribution equality test with an appropriate adjustment for the different (and small) sample sizes we do not reject the null of equal distribution only for the Swedish samples. For the US samples, we fail to obtain strong evidence in favor of the equal distribution assumption. The test statistic here depends on the grouping used for education levels but no grouping leads to strong evidence of equality and for the groupings we used, the range of p-values suggests a strong to marginal rejection of the null. The equal distribution assumption is likely violated for the US education data and we need to study the effects of such violations when considering the properties of our estimator.

[2]We are grateful to a referee for pointing this out to us.

1.2 Empirical Model of Intergenerational Income Mobility

In compliance with the extensive literature on intergenerational income mobility, as a starting point we employ the intergenerational income elasticity to measure the degree to which income status is transmitted from one generation to another. The equation of interest can be written as follows:

$$y_S = \rho y_F + \varepsilon. \tag{1}$$

Here, y_S and y_F are the natural logarithms of permanent incomes of sons and fathers, respectively, and ε is an idiosyncratic error. The parameter of interest, ρ, is referred to as the *intergenerational income elasticity* if the variance of the long-run economic status is different for fathers and sons, i.e., if $Var(y_F) = \sigma_F^2 \neq \sigma_S^2 = Var(y_S)$. In a special case when $\sigma_F^2 = \sigma_S^2$, ρ coincides with intergenerational income correlation. We follow Solon (1992) and use deviations of log income from generation means, so that there is no intercept in equation (1).

The existing literature on intergenerational income mobility has long emphasized that the simplicity of equation (1) should not be taken at the face value. Solon (1999) points out that, though seemingly simple, the basic income mobility equation is still capable of showing that "intergenerational transmission occurs through a multitude of processes" (Solon, 1999, p. 1765). A large empirical body of research maintains that the child's earnings are likely to depend on other aspects of family background, and, thus, this strand of the literature incorporates factors other than parents' earnings into the intergenerational income equation to account for the influence of those factors on intergenerational mobility. Specifically, father's education, occupation, race, union status, industry and country of residence are a few of the characteristics that have been argued to affect intergenerational mobility (see, e.g., Black and Devereux, 2011, for a recent survey of relevant studies).

Furthermore, one of the major concerns with estimating (1), raised in the intro-

duction, is that the standard elasticity of sons' income with respect to that of their fathers' does not appropriately capture the nonlinearity in the intergenerational transmission mechanism of economic status. Therefore, we modify equation (1) to explicitly allow for dependence between the intergenerational elasticity and some characteristic(s) of families. The modified equation can be written as follows:

$$y_S = \rho(\mathbf{z}_1) y_F + \varepsilon, \tag{2}$$

where $\mathbf{z}_1$ contains some family characteristic(s) observable to researchers. While also allowing for explicit variation in intergenerational mobility across the distribution of families, both Bhattacharya and Mazumder (2011) and Murtazashvili (2012) are agnostic about any specific characteristics of the transmission mechanism of earnings from one generation to the next. Similarly, equation (2) permits a flexible functional form describing the intergenerational transmission mechanism of earnings at various points of the distribution of families.

It has been widely discussed in the literature that the permanent earnings of fathers and sons are unobserved. Instead, short-run earnings in the form of either annual earnings or even hourly earnings have to be used as measures of the long-run economic status of fathers and sons. Due to the measurement error in short-run earnings as a proxy for long-run earnings, traditional methods used to estimate intergenerational mobility, which ignore the endogeneity of y_F, suffer from the well-known "attenuation" bias. Traditionally, instrumental variables approaches are used to deal with endogenenity due to measurement error. In the context of intergenerational mobility, Solon (1992) was one of the first to advocate the standard IV approach to estimating intergenerational earnings mobility.

Studies of intergenerational mobility that use IV methods often employ father's characteristics other than income as instrumental variables for fathers' income. In particular, Lefranc and Trannoy (2005) use father's education, occupation, and indicators for living in urban/rural areas as IVs when studying intergenerational earnings mobility in France. Solon (1992) uses father's education to instrument

for father's income when estimating intergenerational income mobility in the USA. Björklund and Jäntti (1997) also employ father's education but add father's occupation dummies as instruments for father's income when comparing intergenerational income mobility in the US and Sweden. Studying the USA, Zimmerman (1992) uses the Duncan index of the prestige of fathers' occupation as an instrument for fathers' earnings.

We follow Björklund and Jäntti (1997) and use fathers' education and fathers' occupation dummies as exogenous variables. Our choice of the exogenous variables is largely driven by data availability. In fact, fathers' education and occupation are the only two characteristics of fathers available in our data set in addition to fathers' earnings. We use fathers' education as $\mathbf{z}_1$ and fathers' occupation as the instrumental variable, $\mathbf{z}_2$, for fathers' earnings. The reason we do not employ fathers' education as $\mathbf{z}_2$ and fathers' occupation as $\mathbf{z}_1$ is that occupational status varies substantially across countries while the level of education is much more standardized. In fact, Table 3 suggests that occupational categories used in the US and Swedish surveys are hard to match. Therefore, our choice of fathers' education as $\mathbf{z}_1$ is meant to facilitate direct cross-country comparisons. It is also easier to justify the assumption of equal distributions of $\mathbf{z}_1$ in the two samples of fathers we are using.

From a theoretical standpoint, parental education along with parental income can be used to explain the differences in financial resources available for investment in children's human capital. Moreover, as Datcher (1982) suggests fathers' education might also reflect fathers' preferences for and value they place on such investment. This idea can be traced back at least to the work by Hill and Stafford (1974) who find a positive association between parents' education and the amount of time they devote to children. Additionally, fathers' years of education used as an exogenous regressor is a way to test "... the idea that broader societal constraints – like discrimination, limits on access to education, and credit market restrictions – limit the earnings prospective of the more successful children from low earnings

backgrounds."(Aydemir, Chen, and Corak, 2009, p. 394).

While our choice of the exogenous variables fits comfortably in the existing theoretical and empirical literature on intergenerational mobility, it is worth mentioning that, when used as instrumental variables, fathers' occupation might also have an independent direct effect on sons' income. However, there is a wide range of literature claiming that it is unlikely to be so. In particular, studies by Sewell and Hauser (1975), Kiker and Condon (1981), Datcher (1982), Corcoran, Gordon, Laren, and Solon (1992), Checchi, Ichino, and Rustichini (1999), and Lefranc and Trannoy (2005) all maintain that fathers' occupational status is correlated with sons' earnings only through its correlation with fathers' income. Furthermore, Solon (1992) shows that even if an instrumental variable has a direct effect on sons' income then, under the typical assumptions, the traditional IV approach will produce an upper bound on the intergenerational income elasticity, which is still a usable result. Thus, we follow this wide literature and employ a full set of fathers' occupational dummies as instrumental variables for fathers' income in our analysis.

A major issue with estimating equation (2) is that, due to the data limitations, we cannot employ conventional econometric methods to estimate the functional coefficient $\rho(\mathbf{z}_1)$. Contrary to the traditional assumption and similar to the intergenerational income mobility studies by Björklund and Jäntti (1997) and Lefranc and Trannoy (2005), we are faced with the situation that information on fathers' and sons' income comes from two different samples.

When equation (1) is the equation of interest, Angrist and Krueger (1992) proposed a parametric method that can deal with the situations that the data come from two samples. In essence, their two-sample IV estimator uses instruments $\mathbf{z}_2$ to predict y_F and then, employs the predicted y_F to estimate ρ. It turns out that a similar nonparametric estimation strategy is possible for the equation in (2). Specifically, under certain conditions, we obtain a consistent nonparametric estimator of $\rho(\mathbf{z}_1)$ using moments obtained from the different data sources. Besides resolving the critical missing data issue, our two-sample nonparametric estimator

remains consistent, provided we have valid instruments, under measurement error in the explanatory variable.

1.3 Two-Sample Nonparametric GMM Estimation

In this section we state a general econometric model of interest, introduce our estimator and discuss its asymptotic properties. We also conduct Monte Carlo simulations to study its properties in samples of the size relevant to the application we consider.

1.3.1 General Statement of Model

In general settings, our model of interest can be written as follows:

$$y_i = \mathbf{x}_i \mathbf{b}(\mathbf{z}_{1i}) + u_i, \tag{3}$$

where y_i is a response variable, $\mathbf{x}_i$ is a $1 \times K$ vector of endogenous explanatory variables, $\mathbf{z}_{1i}$ is a $1 \times L_1$ subvector of the vector of exogenous variables $\mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$, $\mathbf{b}(\cdot)$ is an unknown K-valued function on $\mathbb{R}^{L_1}$, with typical element $b_j(\cdot)$, $j = 1, .., K$, and u_i is an idiosyncratic error. As usual, we assume that $\mathrm{E}[u_i|\mathbf{z}_i] = 0$, i.e. we assume that our instruments $\mathbf{z}_i$ are at least weakly exogenous. We will assume that $L = L_1 + L_2 \geq K$, where $L_2 = \dim(\mathbf{z}_{2i})$, so that there is at least one instrument not in $\mathbf{z}_{1i}$ for every endogenous explanatory variable in $\mathbf{x}_i$.

In the nonparametric literature, model (3) is called a varying (or functional) coefficient model. This model incorporates many linear and partially linear models and it has been used in many applications other than intergenerational income mobility, including exchange rate forecasts (see, e.g., Hong and Lee, 2003) and US unemployment and interest rate analysis (see, e.g., Juhl, 2005). In a parametric context, when $\mathbf{b}(\mathbf{z}_{1i}) \equiv \mathbf{b}_i$ model (3) is called a random coefficient model. Pfeffer-

mann (1984) and Fisher and Voia (2002) generalize the Gauss-Markov theorem to the random coefficient model with missing data. The possibility of correlation between the individual-specific coefficients $\mathbf{b}_i \equiv \mathbf{b}(\mathbf{z}_{1i})$ and $\mathbf{x}_i$ makes (3) a correlated random coefficient model.

When y, $\mathbf{x}$, and $\mathbf{z}$ are contained in a single sample, there are numerous estimation methods that can be used to estimate $\mathbf{b}(\cdot)$ in such models (see, e.g., Cai, Das, Xiong, and Wu, 2006; Zhang, Lee, and Song, 2002). Recently, Cai and Li (2008) proposed a nonparametric GMM approach, which combines the local linear fitting technique and the generalized method of moments (see, e.g., Fan and Gijbels, 1996, for a review of other local smoothing methods). The method is computationally simpler than many other available nonparametric alternatives, so we build on this method in designing our new estimator.

1.3.2 Two-Sample Nonparametric GMM Estimator

For a given point $\mathbf{z}_1 \in \mathbb{R}^{L_1}$ and for $\{\mathbf{z}_{1i}\}$ in a neighborhood of $\mathbf{z}_1$, assuming that $\{b_j(\cdot)\}$ are twice continuously differentiable, we exploit Taylor expansions to approximate $b_j(\mathbf{z}_{1i})$ by a linear function [3] $b_{j0} + \mathbf{b}_{j1}(\mathbf{z}_{1i} - \mathbf{z}_1)$ where $b_{j0} = b_j(\mathbf{z}_1)$ and $\mathbf{b}_{j1} = \frac{\partial b_j(\mathbf{z}_1)}{\partial \mathbf{z}_1}$. So, (3) can be approximated locally by

$$y_i \simeq \mathbf{w}_i \beta + u_i, \tag{4}$$

where $\mathbf{w}_i = \{\mathbf{x}_i, \mathbf{x}_i \otimes (\mathbf{z}_{1i} - \mathbf{z}_1)\}$ is a $1 \times K(1+L_1)$ vector, and $\beta = (b_{10}, ..., b_{K0}, \mathbf{b}'_{11}, ..., \mathbf{b}'_{K1})'$ is a $K(1+L_1) \times 1$ vector of parameters. Thus, for any vector function $\mathbf{Q}(\mathbf{z}_i)$, we can rewrite the standard orthogonality conditions $\mathrm{E}[u_i|\mathbf{z}_i] = 0$ in the following form:

$$\mathrm{E}[\mathbf{Q}(\mathbf{z}_i) u_i | \mathbf{z}_i] = 0. \tag{5}$$

[3] Following Cai and Li (2008), we employ local linear approximation. However, the order of local polynomial approximation can be optimally chosen by cross-validation (see, e.g., Hall and Racine, 2013)

If we observe all the variables y_i, $\mathbf{x}_i$, and $\mathbf{z}_i$ in one sample, the local approximation of conditions (5) will produce the (one-sample) nonparametric GMM (NPGMM) estimator of Cai and Li (2008). We are interested in estimating model (3) using data that come from two samples. To emphasize the distinction between the samples we use superscripts (1) and (2). The difficulty in estimation arises due to the fact that the first data set contains only $\{y_i^{(1)}, \mathbf{z}_i^{(1)}\}$, $i = 1, ..., N_1$, while the second data set contains only $\{\mathbf{x}_l^{(2)}, \mathbf{z}_l^{(2)}\}$, $l = 1, ..., N_2$.

The fundamental difficulty here is that, due to the data structure, calculation of model residuals is infeasible. The dependent variable, the instruments and the independent variables are not available in the same sample and so error-based objective functions, such as the sum of squared residuals, cannot be used. In the setting of nonparametric estimation, this also means that traditional methods of data-driven bandwidth selection and variance estimation are infeasible because they are usually based on residuals. However, under certain conditional moment assumptions it is still possible to obtain a consistent nonparametric estimator based on local averages from the two samples.

In the context of two samples, condition (5) can be approximated by the locally weighted moment condition (6):

$$\mathrm{E}\left[\mathbf{Q}\left(\mathbf{z}_i^{(1)}\right) y_i^{(1)} K_{h_1}\left(\mathbf{z}_{1i}^{(1)} - \mathbf{z}_1\right) - \mathbf{Q}\left(\mathbf{z}_l^{(2)}\right) \mathbf{w}_l^{(2)} \beta K_{h_2}\left(\mathbf{z}_{1l}^{(2)} - \mathbf{z}_1\right)\right] = 0, \qquad (6)$$

where $K_{h_j}(\cdot)$ is a bounded symmetric kernel function on $\mathbb{R}^{L_1}$, $j = 1, 2$, h_1 and h_2 are bandwidths and the dimension of $\mathbf{Q}(\cdot)$ must be at least $K(1 + L_1)$. Though there are many possibilities for $\mathbf{Q}(\cdot)$, we follow Cai and Li (2008) in using the following form

$$\mathbf{Q}(\mathbf{z}_i) = \left(\mathbf{z}_{2i}, \mathbf{z}_{2i} \otimes (\mathbf{z}_{1i} - \mathbf{z}_1)/h_2\right)' : \quad L_2(1 + L_1) \times 1$$

Clearly, for such a choice of $\mathbf{Q}(\mathbf{z}_i)$, a necessary identification condition is $L_2 \geq K$.

Define the following local averages

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{l=1}^{N_2}\mathbf{Q}(\mathbf{z}_l^{(2)})\mathbf{w}_l^{(2)}K_{h_2}(\mathbf{z}_{1l}^{(2)} - \mathbf{z}_1) \tag{7}$$

$$\mathbf{T}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{Q}(\mathbf{z}_i^{(1)})K_{h_1}(\mathbf{z}_{1i}^{(1)} - \mathbf{z}_1)y_i^{(1)} \tag{8}$$

Then, the two-sample nonparametric GMM (TS-NPGMM) estimator we propose has the following simple form

$$\hat{\beta} = (\mathbf{S}_2'\mathbf{S}_2)^{-1}\mathbf{S}_2'\mathbf{T}_1. \tag{9}$$

Implicitly, the estimator in (9), as well as its components $\mathbf{S}_2$ and $\mathbf{T}_1$, are functions of $\mathbf{z}_1$. In essence it is a nonparametric estimator of $\mathbf{b}(\mathbf{z}_1)$ and of its first-order derivatives $\nabla b_j(\mathbf{z}_1)$, where $j = 1, ..., K$, obtained by the GMM for the local neighborhood of $\mathbf{z}_1$.

The following regularity conditions are sufficient for consistency and asymptotic normality of the estimator in (9).

Assumptions:

1. $\{y_i^{(1)}, \mathbf{x}_i^{(1)}, \mathbf{z}_i^{(1)}, u_i^{(1)}\}$ and $\{y_i^{(2)}, \mathbf{x}_i^{(2)}, \mathbf{z}_i^{(2)}, u_i^{(2)}\}$ are two independent samples from the same population, observations are independent across i and only $\{y_i^{(1)}, \mathbf{z}_i^{(1)}\}$ and $\{\mathbf{x}_i^{(2)}, \mathbf{z}_i^{(2)}\}$ are observed. Further, $\mathrm{E}||\mathbf{z}_2^{(j)\prime}\mathbf{x}^{(j)}||^2 < \infty$, $\mathrm{E}||\mathbf{z}_2^{(j)\prime}\mathbf{z}_2^{(j)}||^2 < \infty$, and $\mathrm{E}|u^{(j)}|^2 < \infty$, where $||\mathbf{A}||^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}')$, and $j = 1, 2$.

2. For each $\mathbf{z}_1$, $f(\mathbf{z}_1) > 0$, where $f(\mathbf{z}_1)$ is the density function of $\mathbf{z}_1$, and $\mathbf{b}(\mathbf{z}_1)$ and $f(\mathbf{z}_1)$ are both twice continuously differentiable at any $\mathbf{z}_1 \in \mathbb{R}^{L_1}$.

3. The kernel $K(\cdot)$ is a symmetric, non-negative and bounded second-order kernel function having a compact support; $h_j \to 0$, $h_2/h_1 \to 1$ and $N_j h_j^{L_1} \to \infty$ as $N_j \to \infty$, $j = 1, 2$. Also, $\lim_{N_2\to\infty}\frac{N_2}{N_1} = k$ for some constant, k.

4. **A.** $\mathrm{E}(u|\mathbf{z}) = 0$ and $\mathrm{E}[\pi(\mathbf{z})\pi(\mathbf{z})'|\mathbf{z}_1]$ has a full rank for all $\mathbf{z}_1$, where $\pi(\mathbf{z}) = \mathrm{E}[\mathbf{x}'|\mathbf{z}]$.

B. $\mathrm{E}[\mathbf{x}^{(1)}|\mathbf{z}] = \mathrm{E}[\mathbf{x}^{(2)}|\mathbf{z}] = \mathrm{E}[\mathbf{x}|\mathbf{z}]$.

C. The density of $\mathbf{z}_1$ is identical for both samples and equal to $f(\mathbf{z}_1)$.

Assumptions 1-3 are similar to those in Cai and Li (2008) except for the modifications due to two samples. An important difference is that now we require that the bandwidths used in the two samples satisfy a condition that ensures that no extra terms appear in the asymptotic bias. Deviations from the independence assumption on the error terms and from a homoskedasticity assumption (should we wish to impose it) can be adjusted for using standard methods once we have a consistent estimate of the functional parameter. Assumption 4 is necessary and sufficient for model identification – it makes sure NPGMM works in the two-sample setting. This assumption is similar to the one used by Angrist and Krueger (1992). Their two-sample IV estimator is a parametric two-sample estimator based on equality of unconditional expectations for the two samples while our nonparametric estimator is based on equality of conditional expectations in the context of the two-sample data structure.

We assume that $\mathbf{z}_1$ has the same support in the two samples. Assumption 4 implies that the density of $\mathbf{z}_1$ for the two samples is identical and equal to $f(\mathbf{z}_1)$, which is an even stronger assumption than identical support. If the densities are different, this may cause problems for consistency and nonparametric identification of $\mathbf{b}(\mathbf{z}_1)$. This point is important in intergenerational mobility applications because the two samples are one generation apart. For example, our PSID samples show evidence of different distribution of $\mathbf{z}_1$ for actual fathers and pseudo-fathers. It is therefore important to consider robustness of $\hat{\beta}$ to deviations from this assumption. We return to this point in Section 1.3.3.

A key part of Assumption 4 is that the conditional expectation function for $\mathbf{x}$ given $\mathbf{z}$ is the same for the two samples. If we could observe $\mathbf{x}$ in both samples then given a value of $\mathbf{z} = \mathbf{z}^{(1)} = \mathbf{z}^{(2)}$, the moment condition $\mathrm{E}[\mathbf{x}^{(1)}|\mathbf{z}^{(1)}] = \mathrm{E}[\mathbf{x}^{(2)}|\mathbf{z}^{(2)}] = \mathrm{E}[\mathbf{x}|\mathbf{z}]$ must hold. Under this assumption, sample equivalents of the quantities contained in moment conditions (6) have the same probability limit as for the

one-sample analogue. Intuitively, this guarantees that the probability limit of our estimator is the same as that of Cai and Li (2008).

We provide a formal proof of this intuition in Appendix 1.6 and state only the main result here. Let $\mathbf{H}_j = \text{diag}\{\mathbf{I}_K, h_j\mathbf{I}_{KL_1}\}$, $j = 1, 2$, where $\mathbf{I}_m$ is a $m \times m$ identity matrix.

Theorem 1.1 Under Assumptions 1–4, the TS-NPGMM estimator $\hat{\beta}$ is consistent and asymptotically normal and

$$\sqrt{N_2 h_2^{L_1}} \left[\mathbf{H}_2(\hat{\beta} - \beta) - \frac{h_2^2}{2} \begin{pmatrix} \mathbf{B}_b(\mathbf{z}_1) \\ \mathbf{0} \end{pmatrix} + o_p(h_2^2) \right] \xrightarrow{d} \text{N}(\mathbf{0}, \mathbf{\Psi}), \tag{10}$$

where matrices $\mathbf{\Psi}$ and $\mathbf{B}_b(\mathbf{z}_1)$ are given in Appendix 1.6.

The theorem establishes consistency and asymptotic normality of our estimator and provides its asymptotic variance matrix. It is worth noting that the theorem uses sample-specific quantities N_2, $\mathbf{H}_2$, $\mathbf{T}_1$ and $\mathbf{S}_2$, which distinguish it from its single-sample analogue. The structure of the asymptotic variance matrix (given in Appendix 1.6) is similar to the single-sample case, but the central element of $\mathbf{\Psi}$ is the limiting covariance matrix of $\sqrt{N_2 h_2^{L_1}}(\mathbf{T}_1 - \mathbf{S}_2\beta)$, which is obtained using sample-specific moments so its relation to the single-sample counterpart cannot be established without further assumptions. Specifically, whether the single-sample estimator is relatively more efficient than the two-sample estimator will likely depend on how higher-order moments of $\mathbf{S}_j$ and $\mathbf{T}_j$ compare for the two samples.

In a parametric setting, Inoue and Solon (2010) show that the two-stage least squares version of the two-sample estimator of Angrist and Krueger (1992) is more efficient that the IV version (when both 2SLS and IV are possible). Similarly, in our case, estimator (9) may not deliver the minimal asymptotic variance due to the suboptimal choice of $\mathbf{Q}(\mathbf{z})$. While finding the optimal instruments may be feasible, we prefer the computationally simple form of $\mathbf{Q}(\mathbf{z})$ from Cai and Li (2008). This form is also preferred because it results in a relatively simple form of the asymptotic variance matrix $\mathbf{\Psi}$, whose estimation is sufficiently complicated

as is, due to the two-sample data structure. We address the issue of variance estimation, along with the choice of bandwidth, in Appendix 1.7.

1.3.3 Monte Carlo Simulations

In this section, we conduct Monte Carlo simulations to study the behavior of the TS-NPGMM estimator in realistic non-asymptotic settings relevant for our empirical task. First, we provide graphical illustrations that the TS-NPGMM approach successfully uncovers the true functional forms. Second, we present numerical results showing the rate of convergence as a function of the sample size. Finally, we consider robustness of TS-NPGMM to deviations from the assumption that the distributions of z_1 in the two samples are identical.

We use the sample sizes encountered in our SLLS and PSID data sets and for the robustness study we consider the case where z_1 in both samples is multinomial but there is a slight difference in probabilities. We use the multinomial distribution because fathers' education can be viewed as a discrete random variable for which the multinomial probabilities are equal to the fractions of different education levels in the sample. Correspondingly, in our simulations we allow for deviations from the equal proportions assumption which are equal to the differences in fractions observed in the two US samples.

1.3.3.1 Graphical Illustrations

We start by considering the following data generating process:

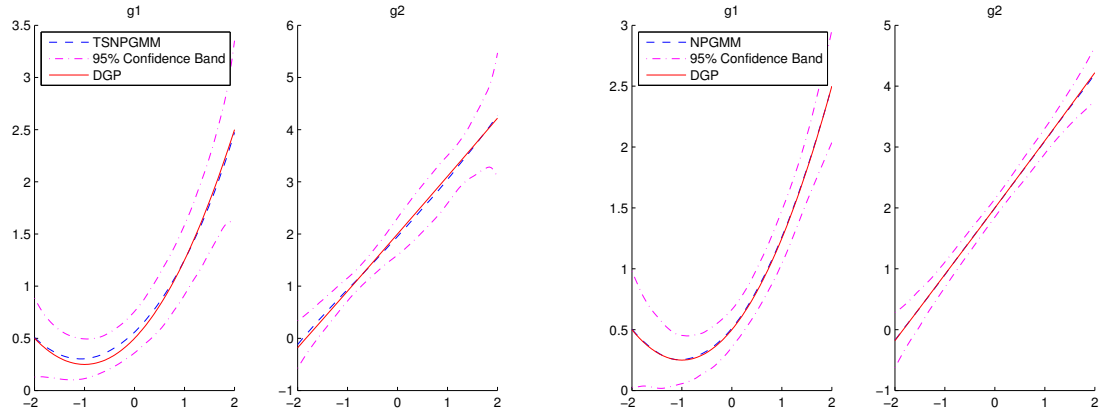
$$Y_i = (0.5 + 0.25{U_i^c}^2 + 0.5U_i^c) + (1 + e^{0.1U_i^c} + U_i^c)X_i + s\epsilon_i, \tag{11}$$

where $U_i^c \sim N(0,1)$ truncated at ± 2, $X_i = (Z_i + \tau\epsilon_i)/\sqrt{1+\tau^2}$, $\epsilon_i \sim N(0,1)$, and $(Z_i, \epsilon_i)' \sim N(0, I_2)$. This is one of the DGPs considered by Su, Murtazashvili, and Ullah (2013). Similarly, we use τ to control the degree of endogeneity and choose s to ensure the signal-noise ratio is 1 when we generate observations on Y_i.

$\{Y_i, U_i, Z_i\}_{i=1}^{N_1}$ and $\{X_j, U_j, Z_j\}_{j=1}^{N_2}$ are two independent samples drawn from a

Figure 1.1: Monte Carlo Simulations for $g_1(u)$ and $g_2(u)$ with 500 Replications

(a) TS-NPGMM ($N_1 = 3,000$, $N_2 = 3,000$) (b) NPGMM (sample size N= 3,000)



population, subject to (11). We do not observe X_i in the first sample and Y_j in the second sample. As a benchmark, we also consider a hypothetical setting where we can observe $\{X_k, Y_k, U_k, Z_k\}_{k=1}^N$ in one sample. We are interested in estimating the two functional coefficients: $g_1(u) = 0.5 + 0.25u^2 + 0.5u$ and $g_2(u) = 1 + e^{0.1u} + u$. For both the two- and one-sample NPGMM estimators, we use the standardized Epanechnikov kernel $k(u) = \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2)\mathbb{I}(|u| \leq \sqrt{5})$ for smoothing and the following simple rule of thumb for bandwidth: $h = s_U\, n^{-1/5}$, where s_U is the standard error of U.

Figure 1.1 provides graphical representations of the two- and one-sample NPGMM approaches based on $N_1 = N_2 = 3,000$ observations and 500 replications. There are at least three interesting observations that can be made from Figure 1.1. First, both of these estimators are remarkably successful in recovering the true functional coefficients. A visual inspection of the two figures reveals an excellent fit. Second, the TS-NPGMM approach appears to have a slightly larger bias than the one-sample NPGMM method. That bias may be caused by a violation of the first equality of Assumption 4B in finite samples. Finally, the confidence bands, calculated using the average variability over replications, are fairly narrow except at the boundaries and the single-sample estimator is substantially more precise than

Table 1.3: Numerical Assessment of TS-NPGMM and NPGMM

Sample Size	TS-NPGMM				NPGMM			
	$\hat{g}_1(u)$		$\hat{g}_2(u)$		$\hat{g}_1(u)$		$\hat{g}_2(u)$	
	MAD	MSE	MAD	MSE	MAD	MSE	MAD	MSE
500	0.445	0.684	0.680	2.275	0.217	0.092	0.229	0.103
800	0.349	0.318	0.496	0.789	0.174	0.057	0.178	0.061
1,000	0.320	0.273	0.441	0.617	0.159	0.046	0.164	0.052
1,500	0.262	0.177	0.363	0.396	0.134	0.033	0.134	0.034
2,000	0.227	0.126	0.323	0.308	0.116	0.025	0.120	0.027
2,500	0.203	0.099	0.287	0.231	0.106	0.020	0.109	0.022
3,000	0.194	0.092	0.271	0.203	0.098	0.017	0.101	0.018
3,500	0.173	0.067	0.245	0.156	0.091	0.015	0.094	0.016
4,000	0.166	0.064	0.236	0.147	0.087	0.013	0.089	0.014

the two-sample estimator. The boundary effect is standard for nonparametric estimators but the visible relative efficiency loss of the two-sample estimator is a new observation. It can be interpreted as the cost of obtaining a feasible estimator.

1.3.3.2 Numerical Assessment

Next, we consider the behavior of TS-NPGMM and NPGMM as the sample size increases. We do so using a grid of $S = 25$ equally spaced points on the interval $[-2, 2]$. We evaluate the estimates of $g_1(u)$ and $g_2(u)$ on the grid and calculate the mean absolute deviation (MAD) and mean squared error (MSE) for each estimator as follows:

$$\mathbf{MAD}_k = \frac{1}{S\,R}\sum_{r=1}^{R}\sum_{s=1}^{S}|\hat{g_k}^{(r)}(u_s) - g_k(u_s)|,$$
$$\mathbf{MSE}_k = \frac{1}{S\,R}\sum_{r=1}^{R}\sum_{s=1}^{S}[\hat{g_k}^{(r)}(u_s) - g_k(u_s)]^2,$$

where $\hat{g_k}^{(r)}(u_s)$, $k = 1, 2$, is an estimate of $g_k(u_s)$ evaluated at grid point s in the r-th replication. This is done over $R = 500$ replications.

We consider nine sample sizes for $N_1 = N_2$. We also looked at unequal samples

but the substantive results are unchanged. We report the corresponding MSE and MAD for TS-NPGMM and NPGMM in Table 1.3. Clearly, as the sample size increases, the MSE and MAD of both estimators decrease quite quickly but remain substantially larger for the two-sample estimator than for their single-sample analogue. This is an interesting result that has to do with the relative bias and relative efficiency of the two-sample versus one-sample estimator.

1.3.3.3 Deviations from Multinomial

Finally, we study robustness of TS-NPGMM to deviations from the assumption of identical distribution of z_1 in both samples. Here we consider the following data generating process:

$$Y_i = (0.0008Z_{1i}^4 - 0.0378Z_{1i}^3 + 0.6017Z_{1i}^2 - 4.0631Z_{1i} + 10.0717)X_i + s\epsilon_i, \quad (12)$$

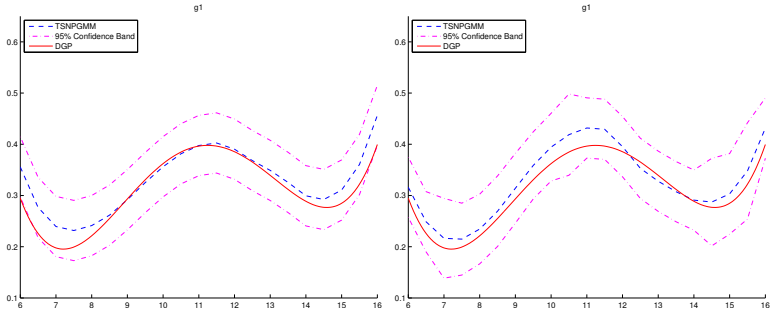
where Z_{1i} is multinomial with K categories, $X_i = (Z_{2i} + \tau\epsilon_i)/\sqrt{1+\tau^2}$, $\epsilon_i \sim N(0,1)$, and $(Z_{2i}, \epsilon_i)' \sim N(0, I_2)$.

To mimic the U.S data, we set the number of observations at 467 for the first sample and 1,613 for the second sample. Also, following the discussion in Section 1.1 we let the K multinomial probabilities in the two samples differ by exactly as much as in the two US samples of fathers. So if we applied the two-sample Chi-square tests of Section 1.1 we would obtain the same p-values, up to a simulation error. The number of replications is 500. The number of categories K ranges between 5 and 9, corresponding to alternative groupings of the PSID education categories.

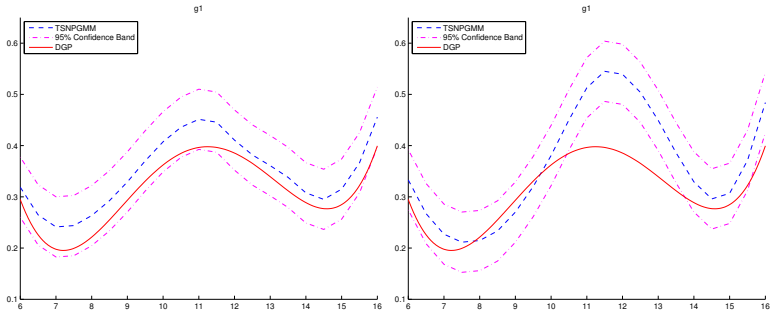
Figure 1.2 presents simulation results for four cases. Panel 1.2(a) shows the performance of TS-NPGMM under the ideal scenario that Z_1 comes from the same multinomial distribution in the two samples. The estimator reveals the true functional form $g(Z_1)$ quite accurately. Panels 1.2(b) to 1.2(d) illustrate the extent of potential bias caused by deviations from the equal distribution assumption, where the deviations are of the magnitude observed in the PSID education data. Clearly

Figure 1.2: Monte Carlo Simulations for Deviations from Same Distribution Assumption

(a) $K = 5$, p-value $= 0.885$ (b) $K = 5$, p-value $= 0.058$



(c) $K = 6$, p-value $= 0.026$ (d) $K = 9$, p-value $= 0.000$



the bias is larger when the differences in the distribution are spread over a larger number of categories K. This corresponds to the range of p-values we obtained when calculating Pearson's Chi-square test for the US education data in Section 1.1. The p-values we report here are the average values over 500 replications.

The main message of Figure 1.2 is that TS-NPGMM shows a degree of robustness to deviations from the assumption of equally distributed Z_1. Extreme deviations captured by near-zero p-values of the Pearson test lead to an upward bias illustrated on Panel 1.2(d). However, for sizable deviations with marginal and strong rejections of the null of equal distributions (p-values of 0.02-0.06), the estimator still correctly estimates the overall functional form, up to a slight upward bias, and contains the true curve within the 95% confidence band, as shown on Panels 1.2(b) and 1.2(c). We will use this robustness feature in the discussion of our empirical findings based on the US data.

Table 1.4: TS-NPGMM Estimates of Intergenerational Income Mobility

Father's Education	USA			Sweden		
	Income Elasticity	Standard Error	Correlation Coefficient	Income Elasticity	Standard Error	Correlation Coefficient
6	0.314	0.023	0.284	0.227	0.026	0.195
9	0.275	0.027	0.249	0.145	0.026	0.124
10	0.371	0.026	0.336	0.207	0.027	0.177
12	**0.422**	**0.024**	**0.382**	**0.296**	**0.023**	**0.254**
14	0.264	0.026	0.239	0.344	0.044	0.295
16	0.321	0.023	0.291	0.478	0.014	0.410

1.4 Empirical Findings and Policy Implications

In this section, we provide our main empirical findings obtained using the TS-NPGMM estimation on the US and Swedish data and we discuss their policy implications. For the estimation, we use the second-order Epanechnikov kernel and a new matching-based method for optimal bandwidth selection and variance estimation, designed to take account of our data structure. We describe the method in Appendix 1.7.

We start by checking the variance equality assumption, which motivates the use of our empirical estimates of $\rho(z_1)$ from equation (2) as intergenerational income correlations rather than elasticities. If the variance of fathers' and sons' earnings is homogeneous, the estimates of $\rho(z_1)$ are equivalent to intergenerational correlations. If the variances are different, the estimates coincide with intergenerational income elasticities. To test the assumption of homogeneity of variances, we employ Levene's (1960) test of equality of variance. Our test statistics reject the null hypothesis of homogeneity for both the US and Swedish samples with p-values of 0.012 and 0.003, respectively. We conclude that the estimates of $\rho(z_1)$ we obtain are indeed the estimates of intergenerational elasticities rather than intergenerational correlations.

Table 1.4 reports the TS-NPGMM estimates, for both the USA and Sweden, of

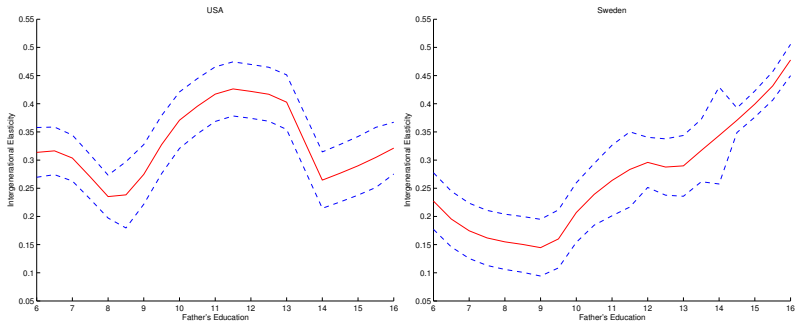
intergenerational income elasticity as a function of fathers' educational attainment for selected values of fathers' education measured in years. In addition, Table 1.4 reports the standard errors of the TS-NPGMM estimates and the estimates of correlation coefficients between sons' and fathers' earnings. We obtain the correlation coefficients by multiplying the US and Swedish intergenerational elasticities by the ratios of the standard deviations in sons' and fathers' earnings that can be found in Table 1.1.

Table 1.4 shows that for the median level of fathers' education, which is at around 12 years (the bold font entries in the table), the corresponding intergenerational income elasticities are about 0.42 and 0.30 for the US and Sweden, respectively. Our TS-NPGMM estimates at the median levels of fathers' education are similar to those reported by Björklund and Jäntti (1997), which are 0.42 and 0.28, respectively. This suggests that income mobility across generations is substantially higher at the median in Sweden than in the US, which is the standard conclusion also found in other studies, involving the US and Scandinavian countries. However, in contrast to the previous studies, our estimates of intergenerational mobility vary a lot outside the median range in both countries.

Similar conclusions regarding intergenerational mobility in the two countries can be made when relying on correlation coefficients rather than intergenerational elasticities. An apparent difference between the estimates of elasticities and correlations is that the latter estimates are noticably smaller than the former ones. This observation is not surprising and it is widely discussed in the literature. For example, Black and Devereux (2011) point out that correlations factor out cross-sectional variations in fathers' and sons' earnings while elasticities might be higher in one society than in another simpy because the variance in sons' earnings is higher in that society. It is for this reason that we report both estimates. Importantly, regardless of which measure of intergenerational mobility we consider, our main conclusions are the same – intergenerational mobility in Sweden is higher than in the USA for the median levels of fathers' education, and it varies a lot outside the

Figure 1.3: TS-NPGMM Estimates of Income Elasticity as a Function of father's Education

(a) US (b) Sweden



Notes: TS-NPGMM estimate, solid line; 95% confidence band, dashed line.

median range in both countries.

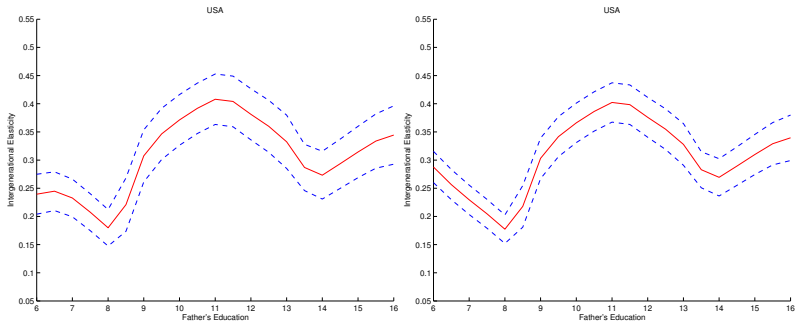
Table 1.5: Robustness checks for U.S income elasticity estimates

Father's	K = 6, p-value = 0.028		K = 5, p-value = 0.078	
Education	Income Elasticity	Standard Error	Income Elasticity	Standard Error
6	0.239	0.018	0.287	0.014
9	0.307	0.023	0.303	0.018
10	0.371	0.022	0.366	0.017
12	0.381	0.023	0.375	0.018
14	0.273	0.021	0.269	0.016
16	0.344	0.026	0.339	0.020

Figure 1.3 presents a visual summary of the intergenerational elasticity estimates reported in Table 1.4, extending them to the entire range of father's educational attainment. Even though a degree of similarity exists in the overall pattern of the two fits – there is an initial trough at 8-9 years followed by a hump at 11-12 years (almost unnoticeable in the case of Sweden) – the profiles of intergenerational income mobility in the two counties look strikingly distinct. Income mobility is lower in the US at the median value of father's education, but the situation is reversed at higher educational attainments. Sweden's income elasticity quickly grows in fathers' education virtually for the entire range of education, starting from at least some high school. American income elasticity decreases no-

Figure 1.4: TS-NPGMM Estimates of Income Mobility for U.S Based on 3 Groupings

(a) K = 6, p-value = 0.028 (b) K = 5, p-value = 0.078



Notes: TS-NPGMM estimate, solid line; 95% confidence band, dashed line.

ticeably when fathers' educational attainment includes the first years of college, and basically stays low (within the error bounds) for higher attainments. So, while Swedish sons seem to inherit the economic status of their well-educated fathers, the US income mobility of sons of educated fathers is actually higher.

When comparing the two parts of Figure 1.3 at education levels below first years of college, we find support for the conventional result that the elasticity is lower in Sweden than in the USA. Also, Figure 1.3(a) clearly shows evidence of a nonlinear relationship between intergenerational income mobility and father's educational attainment for the US. According to Figure 1.3(b), while also nonlinear, this relation is much closer to being linear in Sweden especially when only levels of fathers' education above 9 years are taken into consideration. None of the two patterns provides evidence of a constant income mobility.

Furthermore, the least mobile subpopulation of the Swedish society are sons of highly educated fathers, who also happened to be higher income earners. On the contrary, the least mobile individuals in the US are sons of fathers with a high school degree and lower earnings. Interestingly, in both countries, the income elasticity is lowest when fathers' education is about 8-9 years of schooling, which is close to compulsory education.

In light of the fundings in Section 1.3.3, we carry out robustness checks. We

re-estimate the model for two more groupings of the observed education categories in the US samples, for which the Pearson test of distribution equality leads to a strong or marginal rejection of the null with p-values reported in Section 1.1. Table 1.5 reports the resulting income mobility estimates at selected values of father's education using the grouping with K categories. Figure 1.4 is a visual summary of Table 1.5. Though the new estimates are somewhat lower than in Table 1.4 and Figure 1.3, our conclusions from the comparison with the Swedish estimates are unchanged. The patterns of income elastiticies in Figures 1.3 to 1.4 are remarkably similar in spite of the differences in fractions of various education categories between actual fathers and pseudo-fathers in the PSID samples. The simulation results in Section 1.3.3.3 suggest that TS-NPGMM is robust against this kind of deviations. The empirical results also support that conclusion.

The remarkable observation concerning families with well-educated fathers in the two countries goes against the standard result measured at the median educational attainment of fathers (see, e.g., Björklund and Jäntti, 1997; Österberg, 2000). The median-based estimate seems to mask the widely heterogeneous degree of persistence in earnings across generations with varying family backgrounds. For families with fathers that have at least some college education, intergenerational transmission of earnings turns out stronger in Sweden than in the United States, which is a novel and interesting result.

Given this noteworthy result we are forced to reconsider the link often made between lower income mobility across generations and higher income inequality within a generation across countries (see, e.g., Björklund and Jäntti, 1997; Corak, 2013). Using the Gini coefficient – a popular measure of cross-sectional income inequality – we can make the standard conclusion of the literature on income inequality that Sweden is more equitable than the USA. Indeed, our calculations reported in Table 1.6 show that the Gini coefficient for the USA is larger for both generations than the Gini coefficient for the (roughly) corresponding generations in Sweden. Furthermore, the Gini coefficients in Table 1.6 indicate a (relatively) large

Table 1.6: Gini coefficients for USA and Sweden

	USA	Sweden
Fathers' Earnings	0.318	0.241
Sons' Earnings	0.321	0.188

decrease in income inequality between fathers and sons in Sweden in comparison with a (relatively) small increase in the US. While the magnitudes of the changes we obtain are somewhat different from the ones usually reported in the literature, the direction of the changes are in accordance with the existing findings. Some mismatch in the magnitude is not surprising due to the nature of the data we use for the two countries. In particular, since our samples were constructed using the standard guidelines (see, e.g., Björklund and Jäntti, 1997; Solon, 1992), we inevitably ignore those individuals in both generations who reported zero earnings during the years we consider.

On the one hand, our empirical findings agree with the literature on income inequality in the two countries. On the other, they are in contrast with the existing literature on intergenerational income elasticity. These two observations together highlight a naturally arising possibility of a nonlinear relationship between income mobility across generations and income inequality across countries. In view of our findings, the theoretical conjecture of Solon (2004) that more income equality translates into more intergenerational mobility, known as the "Great Gatsby Curve", may need to be revisited and possibly reversed for important parts of countries' populations.

Do our findings contradict all the existing literature on the intergenerational income transmissions in Sweden and the US? A careful review of previous studies on the subject reveals a surprising conclusion. Several studies have pointed out the possibility of finding the results we find. For example, Peters (1992, p. 466) writes: "[m]uch of the theoretical literature on intergenerational mobility maintains that the interactions between parents' income and ... parents' education ... are

complex," and goes on estimating a model with an unexpected but statistically significant negative effect of fathers' education on sons' earnings, for families with fathers who have at least some college education. The resulting magnitude of the coefficient estimate on the interaction term between fathers' earnings and a dummy variable for whether the father attended some college suggests that there might be no direct relationship between fathers' and sons' earnings for families with fathers who have at least some college education.

Another study indirectly supporting our empirical results is by Aydemir, Chen, and Corak (2009). They document a generational reversal of earnings, that is a situation where sons from below-average backgrounds become above-average earners in their adulthood if the average parental education levels are controlled for. As Peters (1992) speculates, there can be at least two theoretical explanations for the mixed evidence on the interaction effect between parental earnings and education on sons' earnings. First, better educated parents might be more efficient investors. Second, better educated parents might have better access to capital markets and therefore might be able to reduce the dependency of investment on income.

Why do we observe these particular patterns of intergenerational elasticity as a function of fathers' education in the US and Sweden? First, the observed differences can reflect the differences in educational policies targeting low income families in these countries (see, e.g., Bratsberg, Røed, Raaum, Naylor, Jäntti, Eriksson, and Österbacka, 2007). More specifically, the public educational system in Sweden is known to be centralized while it is highly decentralized in the US. Furthermore, the sources of public funding for schools are different in the two countries. In Sweden, schools are funded by the central government while in the USA, local authorities (cities, counties) are heavily involved.

The theoretical framework proposed by Solon (2004) for explaining the differences in intergenerational mobility across countries allows for an explicit effect of government policy in public investment in children's human capital on intergenerational mobility. In particular, Solon's (2004) theoretical model implies that the

more progressive the public spending policy in children's education, the higher the intergenerational mobility in the society. The differences in public policies across countries should be very important due to the role the family background plays in educational decisions. After all, a parental choice to locate in a community is characterized by a combination of such factors as the local tax rates, housing prices and quality of schools.

Interestingly, we observe that the US elasticity is at a minimum for 8-9 and 14 years of fathers education. These two levels of education represent important landmarks in a young man's life in the US, corresponding to the decisions whether to pursue a higher level of education or not. It is possible that the US fathers who once decided not to proceed with their own education have an incentive to persuade their sons to proceed with theirs, while the fathers who achieved higher levels of education and have less personal regrets with respect to their own educational achievements do not have such strong incentives. In this way, the US fathers with 8-9 and 14 years of education can increase their sons' chances of higher earnings in the future.

Given the differences in the educational systems between the USA and Sweden, as well as the differences in public policies in the two countries it is not surprising to find that the family background plays a more important role for US families with low levels of father's education than for Swedish families. The Swedish educational system has been characterized as more successful "in providing all citizens with sufficiently high basic skills so that, particularly at the bottom of parents' earnings distribution, the adult earnings of sons are independent of their parents' economic resources" (Bratsberg at al., 2007, p. C73).

1.5 Conclusion

This paper studies intergenerational income mobility in the US and Sweden while allowing for a flexible nonlinear relationship between fathers' and sons' earnings that has not been previously considered in the literature. Our choice of the country

pair is driven by the previous findings that the Scandinavian countries have the lowest annual income inequality while the US is among the countries with the highest inequality.

In order to estimate the flexible relationship between fathers' and sons' earnings we develop a new nonparametric estimator that addresses the three main concerns arising when conducting cross-country comparisons in intergenerational income transmission – constancy of income mobility across the distribution of families, measurement error in father's long-run economic status, and major missing data issues, which include a two-sample data structure and attrition in later generations. First, we allow for a flexible nonparametric functional form between intergenerational income mobility and observable family background characteristics represented by father's education in our analysis. Second, we exploit an instrumental variable approach to account for measurement error in father's permanent income (using father's occupation as an instrument for father's income). Third, we design a two-sample nonparametric estimator, similar in spirit to the parametric estimator of Angrist and Krueger (1992), to deal with the fact that fathers' and sons' earnings come from different samples.

When we employ our estimator for estimation of intergenerational income mobility in the United States and Sweden we find that the character of inequality in the two countries is strikingly different. Even though the median mobility measures we obtain are similar to those reported in other studies, our mobility estimates for the entire population of families deviate greatly from these median-based levels. Furthermore, our empirical findings suggest that family background captured by father's education matters for both Sweden and the US. For sons of fathers with more than 14 years of education, earnings transmission is stronger in Sweden, with the transmission strength quickly increasing in father's education. For sons whose fathers have educational attainment of less than 14 years of education, earnings transmission is stronger in the US. Most importantly, our empirical results show the patterns characterizing the relationship between intergenerational income mobility

and family background that are useful for designing targeted policy recommendations. Besides permitting detailed cross-country comparisons in income mobility, these results allow to track the effects of various policy changes on specific subgroups of the population. Finally, we advocate future research on the relationship between intergeneratioal mobility and income inequality as our empirical findings suggest a potentially nonlinear nature of this relationship.

1.6 Appendix: Formal Statement and Proof of Theorem

Here we provide a formal proof of consistency and asymptotic normality for the TS-NPGMM estimator. For ease of reference, we adopt the following notation. Let $\mu_2(K) = \int \mathbf{v}\mathbf{v}'K(\mathbf{v})d\mathbf{v}$ and $\mu = \int K^2(\mathbf{v})d\mathbf{v}$. Define

$$\mathbf{R}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} K_{h_j}(\mathbf{z}_{1i}^{(j)} - \mathbf{z}_1)\mathbf{Q}(\mathbf{z}_i^{(j)})\sum_{k=1}^{K} R_k(\mathbf{z}_{1i}^{(j)}, \mathbf{z}_1)x_{ik}^{(j)}, \tag{13}$$

where $R_k(\mathbf{z}_{1i}, \mathbf{z}_1) = b_k(\mathbf{z}_{1i}) - b_{k0} - \mathbf{b}_{k1}(\mathbf{z}_{1i} - \mathbf{z}_1) - \frac{1}{2}(\mathbf{z}_{1i} - \mathbf{z}_1)'\frac{\partial^2 b_k(\mathbf{z}_1)}{\partial \mathbf{z}_1^2}(\mathbf{z}_{1i} - \mathbf{z}_1)$,

$$\mathbf{B}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} K_{h_j}(\mathbf{z}_{1i}^{(j)} - \mathbf{z}_1)\mathbf{Q}(\mathbf{z}_i^{(j)})\frac{1}{2}\sum_{k=1}^{K}(\mathbf{z}_{1i}^{(j)} - \mathbf{z}_1)'\frac{\partial^2 b_k(\mathbf{z}_1^{(j)})}{\partial \mathbf{z}_1^2}(\mathbf{z}_{1i}^{(j)} - \mathbf{z}_1)x_{ik}^{(j)}, \tag{14}$$

and

$$\mathbf{T}_j^* = \frac{1}{N_j}\sum_{i=1}^{N_j} K_{h_j}(\mathbf{z}_{1i}^{(j)} - \mathbf{z}_1)\mathbf{Q}(\mathbf{z}_i^{(j)})u_i^{(j)}, \tag{15}$$

where $j = 1, 2$. Denote the first-sample analogue of $\mathbf{S}_2$ by $\mathbf{S}_1$. Clearly, $\mathbf{S}_1$, $\mathbf{R}_1$, $\mathbf{B}_1$ and $\mathbf{T}_j^*$ are not feasible because $\mathbf{x}_i^{(1)}$, $y_i^{(2)}$, and $u_i^{(j)}$, $j = 1, 2$, are not observed. However, it turns out that for the asymptotic results to apply, we will need assumptions on these quantities.

Theorem. Under Assumptions 1–4, we have

$$\sqrt{N_2 h_2^{L_1}}\left[\mathbf{H}_2(\hat{\beta} - \beta) - \frac{h_2^2}{2}\begin{pmatrix}\mathbf{B}_b(\mathbf{z}_1)\\ \mathbf{0}\end{pmatrix} + o_p(h_2^2)\right] \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{\Psi}), \tag{16}$$

where $\mathbf{\Psi} = f^{-2}(\mathbf{z}_1)(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{\Phi}\mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}$ with $\mathbf{\Omega} = \mathbf{\Omega}(\mathbf{z_1}) = \mathbf{E}(\mathbf{z_2'x}|\mathbf{z_1})$ and $\mathbf{S} = \mathbf{S}(\mathbf{z}_1) = \mathrm{diag}\{\mathbf{\Omega}, \mathbf{\Omega} \otimes \mu_2(K)\}$ and $\mathbf{\Phi}$ being the limiting covariance matrix of $\sqrt{N_2 h_2^{L_1}}(\mathbf{T}_1 - \mathbf{S}_2\beta)$. In addition, $\mathbf{B}_b(\mathbf{z}_1) = \int \mathbf{D}(\mathbf{v}, \mathbf{z}_1)K(\mathbf{v})d\mathbf{v} = (\mathrm{tr}(\nabla^2 b_j(\mathbf{z}_1)\mu_2(K)))$,

$$\mathbf{D}(\mathbf{v}, \mathbf{z}_1) = \begin{pmatrix} \mathbf{v}'\nabla^2 b_1(\mathbf{z}_1)\mathbf{v} \\ \vdots \\ \mathbf{v}'\nabla^2 b_K(\mathbf{z}_1)\mathbf{v} \end{pmatrix}, \text{ and } \nabla^2 b_j(\mathbf{z}_1) = \frac{\partial^2 b_j(\mathbf{z}_1)}{\partial \mathbf{z}_1 \partial \mathbf{z}_1'}.$$

Proof of Theorem: First, notice $\hat{\beta} - \beta = (\mathbf{S}_2'\mathbf{S}_2)^{-1}\mathbf{S}_2'(\mathbf{T}_1 - \mathbf{S}_2\beta)$. Then, $\mathbf{H}_2\hat{\beta} = \mathbf{H}_2(\mathbf{S}_2'\mathbf{S}_2)^{-1}\mathbf{H}_2\mathbf{H}_2^{-1}\mathbf{S}_2'\mathbf{T}_1 = (\tilde{\mathbf{S}}_2'\tilde{\mathbf{S}}_2)^{-1}\tilde{\mathbf{S}}_2'\mathbf{T}_1$, where $\tilde{\mathbf{S}}_2 = \mathbf{S}_2\mathbf{H}_2^{-1}$, and we can express $\mathbf{T}_1$ as $\mathbf{T}_1 = \mathbf{S}_1\beta + \mathbf{R}_1 + \mathbf{B}_1 + \mathbf{T}_1^*$. Then, $\mathbf{T}_1 - \mathbf{S}_2\beta = (\mathbf{S}_1 - \mathbf{S}_2)\beta + \mathbf{R}_1 + \mathbf{B}_1 + \mathbf{T}_1^*$, and we can write

$$\mathbf{H}_2(\hat{\beta} - \beta) - (\tilde{\mathbf{S}}_2'\tilde{\mathbf{S}}_2)^{-1}\tilde{\mathbf{S}}_2'[(\tilde{\mathbf{S}}_1\mathbf{H}_1 - \tilde{\mathbf{S}}_2\mathbf{H}_2)\beta + \mathbf{R}_1 + \mathbf{B}_1] = (\tilde{\mathbf{S}}_2'\tilde{\mathbf{S}}_2)^{-1}\tilde{\mathbf{S}}_2'\mathbf{T}_1^*. \quad (17)$$

The proof of Theorem 2 from Cai and Li (2008) shows that $\sqrt{N_j h_j^{L_1}}\mathbf{T}_j^*$, where $j = 1, 2$, is asymptotically normal with zero mean and finite variance. Further, by Proposition 1 of Cai and Li (2008), $\mathbf{B}_j = O_p(h_j^2)$ and $\mathbf{R}_j = o_p(h_j^2)$, where $j = 1, 2$. These results imply that the last term on the left-hand side of (17), which contains $\mathbf{B}_1$, contributes to the asymptotic bias, while the term containing $\mathbf{R}_1$ is negligible in probability. Condition $h_1/h_2 \to 1$ of Assumption 3 ensures that $\mathbf{R}_1 = o_p(h_2^2)$ and $\mathbf{B}_1 = O_p(h_2^2)$. Then, to establish consistency of (9), we are left with determining the behavior of $(\tilde{\mathbf{S}}_1\mathbf{H}_1 - \tilde{\mathbf{S}}_2\mathbf{H}_2)$. Notice that, by Proposition 1 of Cai and Li (2008), $\tilde{\mathbf{S}}_1 - \tilde{\mathbf{S}}_2 = o_p(1)$ and

$$\sqrt{N_2 h_2^{L_1}}(\tilde{\mathbf{S}}_1\mathbf{H}_1 - \tilde{\mathbf{S}}_2\mathbf{H}_2) =$$

$$\sqrt{N_2 h_2^{L_1}}(\tilde{\mathbf{S}}_1 - f(\mathbf{z}_1)\mathbf{S})\mathbf{H}_1 - \sqrt{N_2 h_2^{L_1}}(\tilde{\mathbf{S}}_2 - f(\mathbf{z}_1)\mathbf{S})\mathbf{H}_2 - \sqrt{N_2 h_2^{L_1}} f(\mathbf{z}_1)\mathbf{S}(\mathbf{H}_2 - \mathbf{H}_1), \quad (18)$$

where $\tilde{\mathbf{S}}_j \to f(\mathbf{z}_1)\mathbf{S}$ for both samples due to Proposition 1 of Cai and Li (2008) and Assumption 4. Condition $h_2/h_1 \to 1$ of Assumption 3 guarantees that the last term of (18) is negligible in probability. Condition $\lim_{N_2 \to \infty} \frac{N_2}{N_1} = k$ of Assumption 3 allows to rewrite the first term of (18) as $\sqrt{k}\sqrt{N_1 h_1^{L_1}}(\tilde{\mathbf{S}}_1 - f(\mathbf{z}_1)\mathbf{S})\mathbf{H}_1$. Then, the first two terms of (18) are also negligible in probability due to Proposition 1 of

Cai and Li (2008). Thus, the consistency of TS-NPGMM is established, and the order of the bias term in expression (17) is h_2^2.

Second, observe that $\sqrt{N_2 h_2^{L_1}}\mathbf{H}_2(\hat{\beta}-\beta) = \sqrt{N_2 h_2^{L_1}}(\tilde{\mathbf{S}}_2'\tilde{\mathbf{S}}_2)^{-1}\tilde{\mathbf{S}}_2'(\mathbf{T}_1 - \mathbf{S}_2\beta)$. Then,

$$\sqrt{N_2 h_2^{L_1}}\left[\mathbf{H}_2(\hat{\beta}-\beta) - \frac{h_2^2}{2}\begin{pmatrix}\mathbf{B}_b(\mathbf{z}_1)\\ \mathbf{0}\end{pmatrix} + o_p(h_2^2)\right] \xrightarrow{d} \mathrm{N}(\mathbf{0}, (\tilde{\mathbf{S}}_2'\tilde{\mathbf{S}}_2)^{-1}\tilde{\mathbf{S}}_2'\mathbf{\Phi}\tilde{\mathbf{S}}_2(\tilde{\mathbf{S}}_2'\tilde{\mathbf{S}}_2)^{-1}), \tag{19}$$

where $\mathbf{\Phi}$ is the limiting covariance matrix of $\sqrt{N_2 h_2^{L_1}}(\mathbf{T}_1 - \mathbf{S}_2\beta)$. Using Proposition 1 from Cai and Li (2008) the asymptotic variance of the left-hand side of (19) becomes $f^{-2}(\mathbf{z}_1)(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{\Phi}\mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}$, where $\mathbf{S} = \mathbf{S}(\mathbf{z}_1) = \mathrm{diag}\{\mathbf{\Omega}, \mathbf{\Omega} \otimes \mu_2(K)\}$. *QED*.

1.7 Appendix: Bandwidth Selection and Variance Estimation

Bandwidth selection is not straightforward for TS-NPGMM because y_i and $\mathbf{x}_i$ are not contained in the same sample. Therefore, the calculation of residuals is infeasible and such standard data-driven tools of residual-based bandwidth selection as least squares cross-validation cannot be applied directly.

Similarly, the main issue in variance estimation is how to obtain an estimate of the limiting covariance matrix of $\sqrt{N_2 h_2}(\mathbf{T}_1 - \mathbf{S}_2\beta)$ if $\mathbf{T}_1$ and $\mathbf{S}_2$ are not available in one sample.

We propose first obtaining a surrogate sample using the following procedure:

Step 1: We match $\mathbf{x}_j^{(2)}$ with $y_i^{(1)}$ by matching $\mathbf{z}_j^{(2)}$ to $\mathbf{z}_i^{(1)}$. That is, for a given value of $\mathbf{z}_j^{(2)}$, we look for such i that $\mathbf{z}_i^{(1)} = \mathbf{z}_j^{(2)}$. The $y_i^{(1)}$ corresponding to that $\mathbf{z}_i^{(1)}$ is the value of y matched to $\mathbf{x}_j^{(2)}$. Specifically, in our empirical analysis, we match son's income with father's income by choosing equal values of father's education, reported in both samples. Not surprisingly, there are more than one such matching observations for any value of father's education. That is, for each value $\mathbf{x}_j^{(2)}$ we have several matched values of $y_i^{(1)}$.

Step 2: We take an average of the subsample of $y_i^{(1)}$ matched to $\mathbf{x}_j^{(2)}$. This produces a single value of the matched y – we denote it by $\bar{y}_j^{(2)}$ – and a surrogate full sample $\{\bar{y}_j^{(2)}, \mathbf{x}_j^{(2)}, \mathbf{z}_j^{(2)}\}$, $j = 1, \ldots, N_2$.

In essence, this procedure is based on Assumption 4B. It can be shown that, under Assumption 4B, $\mathrm{E}(y^{(1)}|\mathbf{z}) = \mathrm{E}(y^{(2)}|\mathbf{z})$ and so, given $\mathbf{z}$, we expect to observe the same values of y in both samples. Basically, we estimate the value of y for the sample that does not contain it with the average of the matched values of y from the sample where y is actually observed.

Once we have the full sample, we apply the standard leave-one-out cross-validation technique to obtain the optimal bandwidth and use it in both samples.

Similarly, we use the surrogate sample $\{\bar{y}_j^{(2)}, \mathbf{x}_j^{(2)}, \mathbf{z}_j^{(2)}\}$ and our consistent estimator $\hat{\beta}$ to estimate $\mathbf{\Psi}$ by

$$\hat{\mathbf{\Psi}} = \hat{f}^{-2}(\hat{\mathbf{S}}'\hat{\mathbf{S}})^{-1}\hat{\mathbf{S}}'\hat{\mathbf{\Phi}}\hat{\mathbf{S}}(\hat{\mathbf{S}}'\hat{\mathbf{S}})^{-1},$$

where, for scalar $\mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{x}$,

$$\begin{aligned}
\hat{f} &= \frac{1}{N_2 h}\sum_{i=1}^{N_2} K_h\left(\mathbf{z}_{1i}^{(2)} - \mathbf{z}_1\right) \\
\hat{\mathbf{S}} &= \operatorname{diag}\{\hat{\mathbf{\Omega}}, \hat{\mathbf{\Omega}} \otimes \mu_2(K)\} \\
\hat{\mathbf{\Omega}} &= \frac{\sum_{i=1}^{N_2} \mathbf{z}_{2i}^{(2)}\mathbf{x}_i^{(2)} K_h\left(\mathbf{z}_{1i}^{(2)} - \mathbf{z}_1\right)}{\sum_{i=1}^{N_2} K_h\left(\mathbf{z}_{1i}^{(2)} - \mathbf{z}_1\right)} \\
\hat{\mathbf{\Phi}} &= \frac{1}{N_2}\left[\sum_{i=1}^{N_2} \hat{\mathbf{q}}_i\hat{\mathbf{q}}_i' + 2\sum_{i\neq j}\hat{\mathbf{q}}_i\hat{\mathbf{q}}_j'\right] \\
\hat{\mathbf{q}}_i &= \mathbf{Q}\left(\mathbf{z}_i^{(2)}\right) K_h\left(\mathbf{z}_{1i}^{(2)} - \mathbf{z}_1\right)\left[\bar{y}_i^{(2)} - \left(\mathbf{x}_i^{(2)}, \mathbf{x}_i^{(2)}(\mathbf{z}_{1i}^{(2)} - \mathbf{z}_1\right)\hat{\beta}\right]
\end{aligned}$$

Here, $\hat{f}$, $\hat{\mathbf{S}}$ and $\hat{\mathbf{\Omega}}$ are feasible without the surrogate sample and are consistent for the population equivalents under standard assumptions. However, $\hat{\mathbf{\Phi}}$ would not be feasible without our matching procedure and its consistency depends on an additional assumption of homoskedasticity across samples. Specifically, since the weighted local residuals $\hat{\mathbf{q}}_i$ represent the sample (2) residuals, the estimates of the limiting covariance matrix based on $\hat{\mathbf{q}}_i$ are consistent if the conditional (co)variance of the error terms in the selected samples is the same, i.e., if $\mathrm{E}\left[\left(u_i^{(1)}\right)^2|\mathbf{z}\right] = \mathrm{E}\left[\left(u_i^{(2)}\right)^2|\mathbf{z}\right] = \mathrm{E}\left[u_i^2|\mathbf{z}\right]$ and $\mathrm{E}\left[u_i^{(1)}u_j^{(1)}|\mathbf{z}\right] = \mathrm{E}\left[u_i^{(2)}u_j^{(2)}|\mathbf{z}\right] = \mathrm{E}\left[u_i u_j|\mathbf{z}\right]$. In other words, we have to assume homoskedasticity across samples whether or not we assume homoskedasticity and uncorrelatedness across i within each sample. In essence $\hat{\mathbf{\Psi}}$ is a version of heteroskedasticity and autocorrelation robust variance estimator based on the surrogate sample.

A further note on homoskedasticity is in order. Our basic consistency and asymptotic normality result in Theorem 1 holds under unrestricted conditional er-

ror variance $\mathrm{E}\left[\left(u_i^{(j)}\right)^2 |\mathbf{z}\right]$, $j = 1, 2$ (provided it is finite – Assumption 1). That is, in principle the errors are allowed to be heteroskedastic across observations within a sample *and* across samples. Even the independence of $u_i^{(j)}$ across i, implied by the first part of Assumption 1, can be relaxed with no consequence for the result as the asymptotic variance matrix $\mathbf{\Psi}$ would reflect the non-equal conditional covariances. The product and cross-product terms $\hat{\mathbf{q}}_i\hat{\mathbf{q}}_j'$ in our estimator of $\mathbf{\Psi}$ provide for the required adjustment to handle heteroskedasticity and autocorrelation within the surrogate sample, but not across the samples.

Chapter 2

Sparse Sieve MLE

The Dantzig selector (DS) was recently introduced to deal with linear regressions in which the number of parameters is very large, possibly larger than the number of observations, but some parameters are believed to be zero – a setting known as a sparsity scenario (Candes and Tao, 2007). DS is attractive because of its property – known as the oracle inequality – to achieve a loss very similar to what we would get if we were told (by an oracle) which elements of the true parameter vector are zero (see, e.g., Koltchinskii, 2009). Unlike the LASSO estimator, which shares similar oracle properties, DS gives parameter estimates with the smallest l_1 norm and is computationally simpler because it reduces to a linear programming problem (see, e.g., Bickel, Ritov, and Tsybakov, 2009).

In this paper we consider using DS in a semiparametric sieve maximum likelihood estimation (SMLE) under a sparsity scenario. Basically, we employ DS in an adaptive nonparametric copula density estimation where the number of sieve farameters is potentially larger than the sample size but the sieve parameter space is sparse. Therefore, this work is related to the sparse density estimation via l_1 penalization (SPADES) of Bunea, Tsybakov, Wegkamp, and Barbu (2010), who consider a LASSO-type penalized objective function. Instead, we use the DS approach, minimizing the l_1 norm of the parameter vector directly.

The goal is to use the nonasymptotic nature of the oracle inequalities to achieve

in finite samples what SMLE achieves only asymptotically – an estimator that dominates the conventional, independence-based QMLE. In other words, the primary purpose of using DS here is relative efficiency and improved finite sample properties, not model selection.

The rest of this chapter is organized as follows. Section 2.1 presents our estimator: the Dantzig selector based Sieve MLE (DS-SMLE). In Section 2.2, we illustrate the finite sample performance of DS-SMLE through simulations. Section 2.3 is an application of DS-SMLE to insurance. Section 2.4 concludes.

2.1 Copula-Based SMLE of Parameters in Marginals

2.1.1 SMLE and QMLE

Consider the setting of a panel with T time periods and N individuals. Assume T is fixed and $N \to \infty$. We will fix $T = 2$ for simplicity. Suppose that for each cross section, we have a correctly specified parametric likelihood-based model and we can estimate this model consistently using only the cross sectional data. However, it is usually possible to use the entire panel to obtain more efficient estimators (see, e.g., Amsler, Prokhorov, and Schmidt, 2014; Prokhorov and Schmidt, 2009b).

The estimator we consider is the sieve MLE (SMLE) (see Chen, 2007, for a review). In essence, this is a maximum likelihood estimator which uses a sieve approximation to the true joint log density. Specifically we follow Panchenko and Prokhorov (2013) and consider a sieve approximation of the copula corresponding to the joint density. In this setting, the SMLE attempts to use information contained in the dependence structure between cross sections.

Let $f(y_{it}; \beta), t = 1, 2$, denote the marginal densities for each cross section, indexed by parameter β. Let $h(y_{i1}, y_{i2}; \beta)$ denote the joint density of $(y_{i1,}, y_{i2})$ and let $c(u_1, u_2)$ denote the copula density, corresponding to $h(y_{i1,}, y_{i2}; \beta)$. We are interested in estimation of β – a parameter vector that collects all unknown parameters from the likelihood-based models for the cross sections. By a well

known result due to Sklar (1959),

$$\ln h(y_{i1,}, y_{i2}; \beta) = \ln f(y_{i1}; \beta) + \ln f(y_{i2}; \beta) + \ln c(F(y_{i1}; \beta), F(y_{i2}; \beta)), \tag{1}$$

where $F(y_{it}; \beta)$ denotes the corresponding marginal cdf's. They may be distinct but we will put this aside for the moment.

The SMLE replaces the last term in (1) with a truncated infinite series representation (a sieve) of the copula log density and then carries out the usual optimization over both β and the parameters of that representation. This produces the sieve MLE estimator $\hat{\beta}$. Panchenko and Prokhorov (2013) derive the semiparametric efficiency bound for estimation of β and show that $\hat{\beta}$ achieves it.

Denote the vector of sieve parameters by γ and the sieve approximator by $\ln c_\gamma$. Then, the SMLE maximizes the approximate joint log likelihood

$$\ln L_\gamma(\beta) = \sum_{i=1}^{N} [\ln f(y_{i1}; \beta) + \ln f(y_{i2}; \beta) + \ln c_\gamma(F(y_{i1}; \beta), F(y_{i2}; \beta))] \tag{2}$$

The fundamental logic of the sieve estimation is that, when the space of functions to be approximated is not too complex and the approximation error goes to zero sufficiently fast, we obtain a $\sqrt{N}$-consistent estimator of β (see, e.g., Shen, 1997; Shen and Wong, 1994).

As an alternative we consider the conventional QMLE estimator which maximizes the quasi-log-likelihood

$$\ln L^Q(\beta) = \sum_{i=1}^{N} [\ln f(y_{i1}; \beta) + \ln f(y_{i2}; \beta)]$$

– identicial to the joint log-likelihood under the assumption of independence between y_{i1} and y_{i2}. It is now well understood that the QMLE is consistent for β but the robust, or "sandwich", version of the variance matrix should be used if there is dependence between the cross sections.

The last term in $\ln L_\gamma(\beta)$ is what distinguishes SMLE from QMLE. We have

assumed that the marginals are correctly specified so the marginal score function – the derivative of $\ln f(y_{it}, \beta)$ with respect to β – is zero mean for both cross sections. Correspondingly, the estimator that maximizes $\ln L_\gamma(\beta)$ requires that the copula score is mean zero while the QMLE requires that it is exactly zero, or equivalently, that the copula is the independence copula $c(u, v) = 1$. That is, unlike QMLE, the SMLE implies that the following first-order condition holds:

$$\sum_{i=1}^{N} \nabla_{(\beta,\gamma)} \ln c_\gamma \left(F(y_{i1}; \beta), F(y_{i2}; \beta) \right) = 0$$

We will use this condition in constructing our new estimator.

2.1.2 Bernstein Polynomial Sieve

Let $[0, 1]^2$ denote the unit cube in $\mathbb{R}^2$. For a distribution function $P_c : [0, 1]^2 \to \mathbb{R}$, a bivariate Bernstein polynomial of order $\mathbf{k} = (k_1, k_2)$ associated with P_c is defined as

$$B_{\mathbf{k},P_c}(\mathbf{u}) = \sum_{j_1=0}^{k_1} \sum_{j_2=0}^{k_2} P_c \left(\frac{j_1}{k_1}, \frac{j_2}{k_2} \right) q_{j_1 k_1}(u_1) q_{j_2 k_2}(u_2) \tag{3}$$

where $\mathbf{u} = (u_1, u_2) \in [0, 1]^2$, $q_{j_s k_s}(u_s) = \binom{k_s}{j_s} u_s^{j_s} (1 - u_s)^{k_s - j_s}$. The polynomial is dense in the space of distribution functions on $[0, 1]^2$ and its order $\mathbf{k}$ controls the smoothness of $B_{\mathbf{k},P_c}$, with a smaller k_s associated with a smoother function along dimension s. Moreover, with the conditions $P_c(0, 1) = P_c(1, 0) = 0$ and $P_c(1, 1) = 1$, $B_{\mathbf{k},P_c}(\mathbf{u})$ is a copula function and is referred to as the Bernstein copula associated with P_c. As $\min\{\mathbf{k}\} \to \infty$, $B_{\mathbf{k},P_c}(\mathbf{u})$ converges to P_c at each continuity point of P_c and if P_c is continuous then the convergence is uniform on the unit cube $[0, 1]^2$ (Sancetta and Satchell, 2004; Zheng, 2011).

The derivative of (3) is the bivariate Bernstein density function

$$\begin{aligned} b_{\mathbf{k},P_c}(\mathbf{u}) &= \frac{\partial^2}{\partial u_1 \partial u_2} B_{\mathbf{k},P_c}(\mathbf{u}) \\ &= \sum_{j_1=1}^{k_1} \sum_{j_2=1}^{k_2} w_{\mathbf{k}}(\mathbf{j}) \prod_{s=1}^{2} \beta(u_s; j_s, k_s - j_s + 1) \end{aligned} \tag{4}$$

where, for $\mathbf{j} = (j_1, j_2)$, $w_{\mathbf{k}}(\mathbf{j}) = \mathbf{\Delta} P_c\left(\frac{j_1-1}{k_1}, \frac{j_2-1}{k_2}\right)$ are weights derived using the forward difference operator $\mathbf{\Delta}$, and $\beta(\cdot; \gamma, \delta)$ denotes the probability density function of the β-distribution with parameters γ and δ.

In order to give a mixing interpretation to $w_{\mathbf{k}}$, let $\text{Cube}(\mathbf{j}, \mathbf{k})$ denote a cube given by $((j_1-1)/k_1, j_1/k_1] \times ((j_2-1)/k_2, j_2/k_2]$ with the convention that if $j_s = 0$ then the interval $((j_s-1)/k_s, j_s/k_s]$ is replaced by the point $\{0\}$. Then, the mixing weights $w_{\mathbf{k}}(\mathbf{j})$ are the probabilities of $\text{Cube}(\mathbf{j}, \mathbf{k})$ under P_c. The Bernstein density function $b_{\mathbf{k},P_c}(\mathbf{u})$ can thus be viewed as a mixture of beta densities, and if P_c is a copula, $b_{\mathbf{k},P_c}(\mathbf{u})$ is itself a copula density.

Alternatively, if we interpret P_c as an empirical copula on $\left[\frac{1}{k_1}, \frac{2}{k_1}, ..., \frac{k_1}{k_1}\right] \times \left[\frac{1}{k_2}, \frac{2}{k_2}, ..., \frac{k_2}{k_2}\right]$ then $b_{\mathbf{k},P_c}(\mathbf{u})$ can be viewed as a smoothed copula histogram using β-densities as smoothing functions.

The Bernstein copula density has several attractive properties as a sieve for the space of copula densities, which makes it preferable to other types of sieve. Being a mixture of (a produce of) β-densities, it assigns no weights outside $[0,1]^2$ and it easily extends to dimensions higher than two. Other sieves known to approximate well smooth functions and densities on $\mathbb{R}$ are often subject to the boundary problem and do not extend easily to multivariate settings (see, e.g., Bouezmarni and Rombouts, 2010; Chen, 2007). The Bernstein sieve is a copula density by construction; at the same time, it does not impose symmetry, contrary to other conventional kernels used in mixture models such as multivariate Gaussian (see, e.g., Burda and Prokhorov, 2013).

Most importantly, as a density corresponding to $B_{\mathbf{k},P_c}(\mathbf{u})$, $b_{\mathbf{k},P_c}(\mathbf{u})$ converges, as $\min\{\mathbf{k}\} \to \infty$, to $p_c(\mathbf{u}) \equiv \frac{\partial^2}{\partial u_1 \partial u_2} P_c(\mathbf{u})$ at every point on $[0,1]^2$ where $p_c(\mathbf{u})$ exists,

and if p_c is continuous and bounded then the convergence is uniform(Lorentz, 1986). Uniform approximation results for the univariate and bivariate Bernstein density estimator can be found in Vitale (1975) and Tenbusch (1994).

In what follows we will assume $P_c(\mathbf{u})$ to be a continuous copula. As a result, we will omit subscript P_c and let $b_{\mathbf{k}}(\mathbf{u})$ simply denote the Bernstein copula density with weights w_j, where $j = 1, \ldots, J$, indexes the set $\{j_1, j_2\}$. Consequently, we can write the copula density as follows

$$b_{\mathbf{k}}(\mathbf{u}) = \sum_{j}^{J} w_j g_j(\mathbf{u}),$$

where $g_j(\mathbf{u}) = \prod_{s=1}^{2} \beta(u_s; j_s, k_s - j_s + 1)$.

2.1.3 SMLE with Dantzig Selector

In practice, the SMLE involves a truncation of the Bernstein polynomial approximation at some large values $\mathbf{k}_N \equiv (k_1^*, k_2^*)$. This means there is a large but finite number – possibly different in each coordinate – of the mixing weights w_j in the Bernstein copula density. Let γ_N contain all such mixing weights. Then, $J = \dim\{\gamma_N\} = k_1^* k_2^*$ and it will grow exponentially as we add dimensions. An important issue in adaptive estimation of such models is how to reduce the dimension of γ_N.

2.1.3.1 Dantzig Selector

The Dantzig selector is an "automatic" mechanism for selecting non-zero parameters in highly parameterized problems. It is "automatic" because we do not need to even set the maximum number of non-zero parameters. So long as there are zero and non-zero elements in the parameter vector, that is, so long as a sparsity scenario applies, the method will pick the non-zero parameters correctly.

The initial application of the Dantzig selector was in linear regressions with more regressors than observations. Suppose we have the following regression model

$y = X\theta + u$, where $\theta \in \mathbb{R}^p$, $u \sim N(0, \sigma^2\mathbb{I})$ and X is a $N \times p$ data matrix with possibly fewer rows than columns, i.e. with $N < p$. Then, the Dantzig selector of Candes and Tao (2007) is the solution to the following problem

$$\min_{\theta} ||\theta||_{l_1} \text{ subject to } ||X'(y - X\theta)||_{l_\infty} \leq \lambda_p \, \sigma, \tag{5}$$

where $||\theta||_{l_1} = \sum_{j=1}^{p} |\theta_j|$ is the l_1-norm of θ, $||Z||_{l_\infty} = \max\{|Z_1|, \ldots, |Z_p|\}$ is the l_∞-norm of any vector $Z \in \mathbb{R}^p$, and λ_p is a positive number – a function of p only. Compared to the usual OLS, the Dantzig selector searches for a θ which has the smallest l_1-norm and, within a fixed tolerance level λ, satisfies the normal equations. Beacuse it produces sparse coefficient estimates, it can be used for model selection. For $\lambda = 0$, it reduces to standard OLS.

It is well known (see, e.g., Bickel, Ritov, and Tsybakov, 2009) that this problem can be viewed as a penalized LS problem, written as follows

$$\min_{\theta} \left\{ \text{SSE}(\theta) + 2\lambda_p \sigma \sum_{j=1}^{p} |\theta_j| \right\}, \tag{6}$$

where $\text{SSE}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - X_i\theta)^2$ and the penalty term grows with complexity of θ as measured by the l_1-norm. So the Dantzig selector solves this problem for a vector having the smallest l_1-norm.

The most attractive theoretical property of the Dantzig selector is that there is a nonasymptotic bound on the error in the estimator of θ that is within a factor of $\log p$ of the error achieved if the true predictors are assumed known. To see this, let $\hat{\theta}$ denote the solution. Candes and Tao (2007) show that under certain conditions on X and under a sparsity scenario (which roughly amounts to an identification condition in this model), the following holds with a large probability,

$$||\hat{\theta} - \theta||_{l_2}^2 \leq \text{const} \cdot \lambda_p^2 \cdot \left(\sigma^2 + \sum_{j=1}^{p} \min\left\{\theta_j^2, \sigma^2\right\} \right), \tag{7}$$

where $||\theta||_{l_2} = \sqrt{\theta'\theta}$ and λ_p^2 is of order $O(\log p)$.

Now consider a standard LS estimator in the situation when we know (from an oracle) which θ_j's are significant (i.e., larger than the noise, $|\theta_j| > \sigma$). In this case, we can set equal to zero all the elements of θ that are smaller than σ in magnitude and let the OLS estimate the significant elements. If, for simplicity, X is assumed to be the identity matrix, then the MSE of the LS estimate of θ will contain terms equal to σ^2 for each significant θ_j and terms equal to θ_j^2's for each insignificant θ_j's (i.e., for the coordinates within the noise level). That is, the MSE of this infeasible estimator can be written as follows

$$\mathbb{MSE}_{\text{OLS}} = \sum_{i=1}^{p} \min\{\theta_j, \sigma^2\}$$

When we relax the assumption that X is identity but still allow the oracle to tell us which subset of θ_j's is right to use in the OLS, the MSE will be different. However, Candes and Tao (2007) show that, under certain assumptions on X, $\mathbb{MSE}_{\text{OLS}}$ can still be viewed as a proxy for the MSE in the more general setting, which has the following natural interpretation

$$\sum_{i=1}^{p} \min\{\theta_j, \sigma^2\} = \min_{S \subset \{1,\dots,p\}} ||\theta - \theta_S||_{l_2}^2 + |S|\sigma^2,$$

where S indexes the set of significant θ_j's, θ_S contains θ_j's if j is in S and 0's otherwise and $|S|$ denotes the number of non-zero elements in S. Of course, the first term of this representation is the squared bias of the ideal estimator and the second is its variance

So the DS nearly achieves the MSE of the ideal estimation, in which an oracle tells us the composition of S. Specifically, the MSE of DS in (7) can be written as follows

$$\mathbb{MSE}_{\text{DS}} \leq \text{const} \cdot \lambda_p^2 \cdot (\sigma^2 + \mathbb{MSE}_{\text{OLS}}).$$

In other words, even though no knowledge of the sparsity scenario was used in

estimating $\hat{\theta}$, the estimation error is proportional to $\log p$ times the error rate achieved if the significant X's were known. So the price we pay for choosing the true predictors by DS is quite small as $\log p$ is not a fast rate. This feature is known as the oracle property of DS.

2.1.3.2 Dantzig Selector for Copula Score

It is not difficult to see that under Gaussian errors the constraint in (7) is a constraint on the score function of the underlying likelihood. So the DS can be equivalently interpreted as looking for a sparse θ close to the peak of the normal likelihood. This observation motivates the estimator we propose.

The Dantzig Selector SMLE (DS-SMLE) we propose is the solution to the following minimization problem

$$\min_{\beta,\gamma_N} ||\gamma_N||_{l_1} \quad \text{subject to} \quad \left|\left| \frac{1}{N}\sum_{i=1}^{N} \nabla_{(\beta,\gamma)} \ln c_{\gamma_N}\left(F(y_{i1};\beta), F(y_{i2};\beta)\right) \right|\right|_{l_\infty} \leq r \quad (8)$$

$$\text{and} \quad \frac{1}{N}\sum_{i=1}^{N} \nabla_{\beta} \ln f(y_{it};\beta) = 0, \ t = 1, 2$$

where $c_\gamma(\mathbf{u}) = b_{\mathbf{k}}(\mathbf{u})$ is the Bernstein copula density, and $\nabla_{(\beta,\gamma)}$ denotes the derivative with respect to (β, γ).

The mean zero conditions on the marginal scores correspond to the assumption of correct specification of the marginals, which is our basic supposition. The copula score with respect to β and γ corresponds to the additional terms in the joint log-likelihood. In the fully parametric setting with a correctly-specified (up to a finite dimensional parameter γ) copula family, this score would be zero mean. In our setting, γ represents a function and $\dim\{\gamma\}$ is potentially greater than the sample size. Essentially, our estimator looks for such a vector (β', γ') for which γ has the smallest l_1-norm and the first order conditions characterizing the MLE solution hold within a fixed tolerance level.

This problem is an example of l_1-norm minimization subject to nonlinear constraints. There are equivalent convex formulations for such problems (see, e.g.,

Candes, 2006). We can rewrite (8) as follows

$$\min_{\beta,\gamma_N,x} \sum_{j=1}^{\dim \gamma_N} x_i \quad \text{subject to} \quad -x \preceq \gamma_N \preceq x \tag{9}$$
$$-r\mathbf{1} \preceq \tfrac{1}{N}\sum_{i=1}^{N} \nabla_{(\beta,\gamma)} \ln c_{\gamma_N}(F(y_{i1};\beta), F(y_{i2};\beta)) \preceq r\mathbf{1}$$
$$\tfrac{1}{N}\sum_{i=1}^{N} \nabla_{\beta} \ln f(y_{it};\beta) = 0, \ t = 1,2$$

where $x = \{x_i\}_{i=1}^{\dim \gamma_N}$, $\mathbf{1}$ denotes a conforming vector of ones and "$\preceq$" represents coordinate-wise comparison of vectors. This will be the preferred formulation in practice because standard convex optimization procedures and fast algorithms are available to compute the solution, which includes $\hat{\beta}$ (see, e.g., Birgé and Massart, 1997; Devroye and Lugosi, 2000).

In order to see the relationship between this estimator and the penalized LS problem (6), note that DS-SMLE can be viewed as a solution to the following penalized MLE problem:

$$\min_{\beta,\gamma} \left\{ -\frac{1}{N} \ln L_\gamma(\beta) + r \sum_{j=1}^{\dim\{\gamma\}} |\gamma_j| \right\}, \tag{10}$$

where $\ln L_\gamma(\beta)$ is the copula-based log-likelihood given in (2), in which the marginals are assumed to be correctly specified. This is, of course, the penalized LS criterion from (6), with SSE replaced by $\ln L$, and the logic of our estimator is in essence the same as that of the conventional Dantzig selector – we are choosing the sparsest vector satisfying the Dantzig constraint implied by the penalized problem.

The choice of $\frac{1}{N} \ln L_\gamma(\beta)$ in (10) is natural if we view our problem as a minimization of the Kullback-Leibler distance between the true density $h(y_1, y_2)$ and the sieve-based density $h_\gamma(y_1, y_2; \beta)$, where $h_\gamma(y_1, y_2; \beta) = f(y_1;\beta) \cdot f(y_2;\beta) \cdot c_\gamma(F(y_1;\beta), F(y_2;\beta))$. Let $\mathbb{KL}(f, g)$ denote the Kullback-Leibler distance between arbitrary densities f and g. Then,

$$\arg\min_{\beta,\gamma} \mathbb{KL}(h, h_\gamma) = \arg\min_{\beta,\gamma} \mathbb{E} \ln \frac{h(y_1, y_2)}{h_\gamma(y_1, y_2; \beta)} = \arg\min_{\beta,\gamma} \left[-\mathbb{E} \ln h_\gamma(y_1, y_2; \beta) \right].$$

The expectation we minimize depends on the unknown h, so instead, we approximate it by its empirical counterpart $-\frac{1}{N}\ln L_\gamma(\beta)$. From this perspective, the problem in (10) can be viewed as a minimization of penalized Kullback-Leibler divergence.

2.1.3.3 Oracle Inequality

In this section we provide an oracle property of our estimator. We compare its risk with that of an infeasible procedure in which an oracle tells us which components of γ are insignificant. We start with a result for the copula parameter γ.

Suppose the marginal distributions are known. Then, the DS problem in (8) reduces to looking for the sparsest vector γ such that $\left|\left|\frac{1}{N}\sum_{i=1}^{N}\nabla_\gamma \ln c_\gamma\left(u_{i1},u_{i2}\right)\right|\right|_{l_\infty} \leq r$, where $u_{ij}=F(y_{ij}), j=1,2$, are obtained using the known marginals. Let $\hat{\gamma}$ denote this solution. The next result gives a bound on the KL divergence of the $\hat{\gamma}$-based copula.

Proposition 2.1. *Let $c_\gamma(\mathbf{u})$ be the Bernstein copula sieve, i.e. $c_\gamma(\mathbf{u})=\gamma'\mathbf{g}(\mathbf{u})$, where $\mathbf{g}(\mathbf{u})=(g_1(\mathbf{u}),\ldots,g_J(\mathbf{u}))'$ and $g_j(\mathbf{u})=\prod_{s=1}^{2}\beta(u_s;j_s,k_s-j_s+1), j=1,\ldots,J$. Let $M_j\equiv||g_j(\mathbf{u})||_{l_\infty}$, $j=1,\ldots,J$. Then, with probability close to one, for all $\gamma\in\mathbb{R}^J$*

$$\mathbb{KL}(c,c_{\hat{\gamma}})\leq \mathbb{KL}(c,c_\gamma)+2r\sum_{j=1}^{J}|\hat{\gamma}_j-\gamma_j| \tag{11}$$

Proof. Let $l_{\gamma i}\equiv \ln c_\gamma(u_{1i},u_{2i})$ and let $J\equiv\dim\{\gamma\}$. By definition of $\hat{\gamma}$,

$$-\frac{1}{N}\sum_{i=1}^{N}l_{\hat{\gamma}i}+r\sum_{j=1}^{J}|\hat{\gamma}_j|\leq -\frac{1}{N}\sum_{i=1}^{N}l_{\gamma i}+r\sum_{j=1}^{J}|\gamma_j|,$$

for any $\gamma\in\mathbb{R}^J$. Thus,

$$\mathbb{KL}(c,c_{\hat{\gamma}})\leq \mathbb{KL}(c,c_\gamma)+\frac{1}{N}\sum_{i=1}^{N}(l_{\hat{\gamma}i}-l_{\gamma i})-\mathbb{E}(l_{\hat{\gamma}i}-l_{\gamma i})+r\sum_{j=1}^{J}|\gamma_j|-r\sum_{j=1}^{J}|\hat{\gamma}_j|$$

Define $\xi_j(\mathbf{u_i}) = \frac{g_j(\mathbf{u}_i)}{c_\gamma(\mathbf{u}_i)}$ and let $D_j = \frac{1}{N}\sum_{i=1}^N \{\xi_j(\mathbf{u}_i) - \mathbb{E}\xi_j(\mathbf{u}_i)\}$. Define the event $\Omega = \bigcap_{j=1}^J \{|D_j| \leq r\}$. By concavity of the log-function,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N (l_{\hat{\gamma}i} - l_{\gamma i}) - \mathbb{E}(l_{\hat{\gamma}i} - l_{\gamma i}) &\leq \frac{1}{N}\sum_{i=1}^N \frac{1}{c_\gamma(\mathbf{u}_i)}\left[c_{\hat{\gamma}}(\mathbf{u}_i) - c_\gamma(\mathbf{u}_i)\right] - \mathbb{E}\frac{1}{c_\gamma(\mathbf{u}_i)}\left[c_{\hat{\gamma}}(\mathbf{u}_i) - c_\gamma(\mathbf{u}_i)\right] \\ &= \sum_{j=1}^J \left(\frac{1}{N}\sum_{i=1}^N \frac{g_j(\mathbf{u}_i)}{c_\gamma(\mathbf{u}_i)} - \mathbb{E}\frac{g_j(\mathbf{u}_i)}{c_\gamma(\mathbf{u}_i)}\right)\left[\hat{\gamma}_j - \gamma_j\right]\end{aligned}$$

Therefore,

$$\mathbb{KL}(c, c_{\hat{\gamma}}) \leq \mathbb{KL}(c, c_\gamma) + \sum_{j=1}^J \left(\frac{1}{N}\sum_{i=1}^N \xi_j(\mathbf{u}_i) - \mathbb{E}\xi_j(\mathbf{u}_i)\right)\left[\hat{\gamma}_j - \gamma_j\right] + r\sum_{j=1}^J |\gamma_j| - r\sum_{j=1}^J |\hat{\gamma}_j|$$

Hence, on the event Ω,

$$\begin{aligned}\mathbb{KL}(c, c_{\hat{\gamma}}) &\leq \mathbb{KL}(c, c_\gamma) + r\sum_{j=1}^J |\hat{\gamma}_j - \gamma_j| + r\sum_{j=1}^J |\gamma_j| - r\sum_{j=1}^J |\hat{\gamma}_j| \\ &\leq \mathbb{KL}(c, c_\gamma) + 2r\sum_{j=1}^J |\hat{\gamma}_j - \gamma_{oj}|\,,\end{aligned}$$

where the last inequality follows by the triangle inequality.

Now by the Hoeffding inequality,

$$\mathbb{P}(\Omega) = \mathbb{P}(\bigcap_{j=1}^J |D_j| \leq r) = 1 - \mathbb{P}(\bigcup_{j=1}^J |D_j| > r) \geq 1 - \sum_{j=1}^J \exp(nr^2/(16M_j^2)) = 1 - \delta.$$

□

2.2 Simulations

In this section we study the finite sample behavior of DS-SMLE as well as discuss issues arising when simulating from the Bernstein copula. Our goal is to compare the behavior of DS-SMLE with QMLE and SMLE, where the QMLE is the conventional estimator based on the independence assumption and the SMLE is

the unpenalized SMLE based on the Bernstein copula. The DS-SMLE reduces to SMLE when $r = 0$.

Numerically, the fundamental difference between SMLE and DS-SMLE is that the SMLE estimates the entire vector γ_N for some large value of J_N, while DS-SMLE shrinks the elements of γ_N toward zero and estimates only the non-zero elements.

2.2.1 Simulating from Bernstein copula

A key issue in simulations is how to generate data from the Bernstein copula. The problem is that the standard way of generating observations from an arbitrary copula, known as the conditional cdf method, is too expensive in the settings of the Bernstein copula. The reason for this is that γ is obtained as the first order difference of parameters in the Bernstein copula cdf. As a result, γ basically contains $\Delta P_c(\frac{j_1}{k_1}, \frac{j_2}{k_2})$ and we have to solve a large system of equations to obtain $P_c(\frac{j_1}{k_1}, \frac{j_2}{k_2})$, where

$$\Delta P_c\left(\frac{j_1}{k_1}, \frac{j_2}{k_2}\right) = P_c\left(\frac{j_1+1}{k_1}, \frac{j_2+1}{k_2}\right) - P_c\left(\frac{j_1+1}{k_1}, \frac{j_2}{k_2}\right) - P_c\left(\frac{j_1}{k_1}, \frac{j_2+1}{k_2}\right) + P_c\left(\frac{j_1}{k_1}, \frac{j_2}{k_2}\right)$$

As an alternative, we use the accept-reject approach (see, e.g., Pfeifer, Strassburger, and Philipps, 2009). To introduce the method, suppose we want to generate data from a distribution F with a pdf $f(x)$, which is a complicated distribution and we do not know how to simulate from it directly. The basic idea of the method is to find another distribution G with a pdf $g(y)$, for which we already have an efficient algorithm to generate data. The key is that this distribution should also be very close to $f(x)$. Specifically, the ratio $f(x)/g(x)$ should be bounded by a positive constant M, i.e. $\sup_x\{f(x)/g(x)\} \leq M$. Then we can apply the following procedure:

1. Generate y from $g(y)$

2. Independently generate u from uniform on [0,1]

3. If $u \leq \frac{f(y)}{Mg(y)}$, then set $x = y$ and use x as a sample from $f(x)$. Otherwise, go back to Step 1.

It can be easily shown that $P(Y \leq y | U \leq \frac{f(y)}{cg(y)}) = F(y)$. Also, note that the expected number of steps required to generate one observation from $f(x)$ is M.

We wish to apply the accept-reject method to the Bernstein copula. We use a multivariate uniform distribution as the reference distribution $G(.)$ with the density function $g(.) = 1$. In this case, $M = \sup_u\{b_{\mathbf{k}}(\mathbf{u})/g(\mathbf{u})\} = \max_{\mathbf{u}\in[0,1]^d}\{b_{\mathbf{k}}(\mathbf{u})\}$. The simulation algorithm is as follows:

1. Generate $(u_1, \ldots, u_d)$ from the multivariate uniform distribution. Here d denotes the number of cross-sections.

2. Independently generate u_{d+1} from uniform on [0,1].

3. if $u_{d+1} \leq \frac{b_{\mathbf{k}}(\mathbf{u})}{M}$, then use $(u_1, \ldots, u_d)$ as an observation from the Bernstein copula. Otherwise, go back to step 1.

It is clear that due to the reference distribution G being uniform, we can actually combine Step 1 and 2 into one step.

2.2.2 Sparse Parameter Path

The tuning parameter r is key to the amount of shrinkage done by the DS. As a first step of the simulation exersice we study the behavior of our estimator of γ over all r.

Our data generating process has exponential marginals with $\mu_1 = \mu_2 = 0.5$ and the Bernstein copula with $J = 16$ (four parameters in each dimension), so in total, there are 18 parameters. However, the γ has only four elements out 16 that are

nonzero as shown in the following matrix.

$$\begin{bmatrix} 0 & 0 & 0 & 0.278 \\ 0 & 0 & 0.212 & 0 \\ 0 & 0.244 & 0 & 0 \\ 0.266 & 0 & 0 & 0 \end{bmatrix}$$

This corresponds to a copula with a high negative dependence. The number of observations is 100.

Figure 2.1 shows the estimated parameter paths for the non-zero elements of γ (colored solid lines) and the insignificant elements (dashed red lines). There are two important observations. First, the DSSMLE can correctly identify the non-zero elements in γ. Second, in the region where the zero γ_j's are actually estimated to be close to zero (the region with small l_1), the non-zero γ_j's are estimate to be smaller than the true values. This suggests that the DS-SMLE over-shrinks γ.

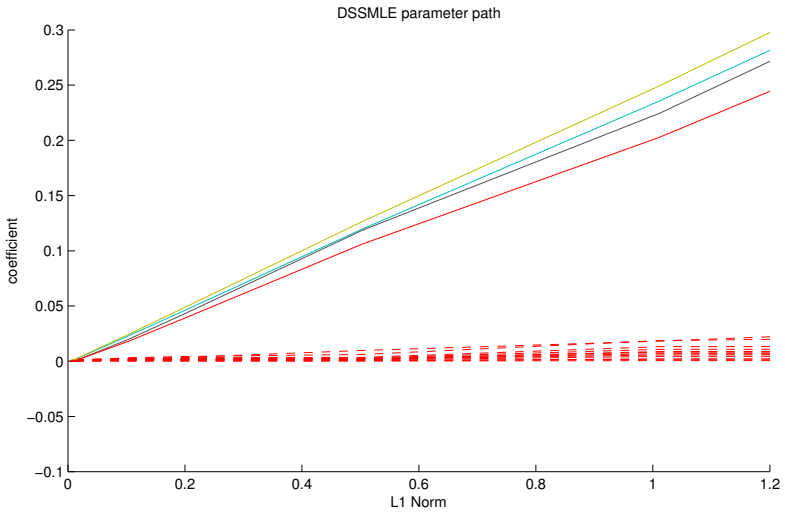
The over-shrinkage result is not uncommon in the DS literature and James and Radchenko (2009) propose a two-step procedure called double Dantzig to overcome this issue. We follow James and Radchenko (2009) and implement the following two-step procedure in our simulations:

1. Run the DS-SMLE using a large value of the tuning parameter. Select the non-zero elements γ_j. Denote the selected set by γ^*.

2. Run the unrestricted SMLE over γ^* and β.

So in effect we run two DS-SMLE where in the second step we set the tuning parameter equal to be zero. A similar procedure called the gaussian Dantzig selector was proposed by (Candes and Tao, 2007, p. 2323) and can be seen as a special case of the double Dantzig of James and Radchenko (2009).

Figure 2.1: DSSMLE Parameter Path



Notes: Plot of estimated coefficients for different values of λ. The solid lines represent the variables which are nonzeros in the true setting of γ. The dashed lines correspond to the remaining variables.

2.2.3 Simulation Results

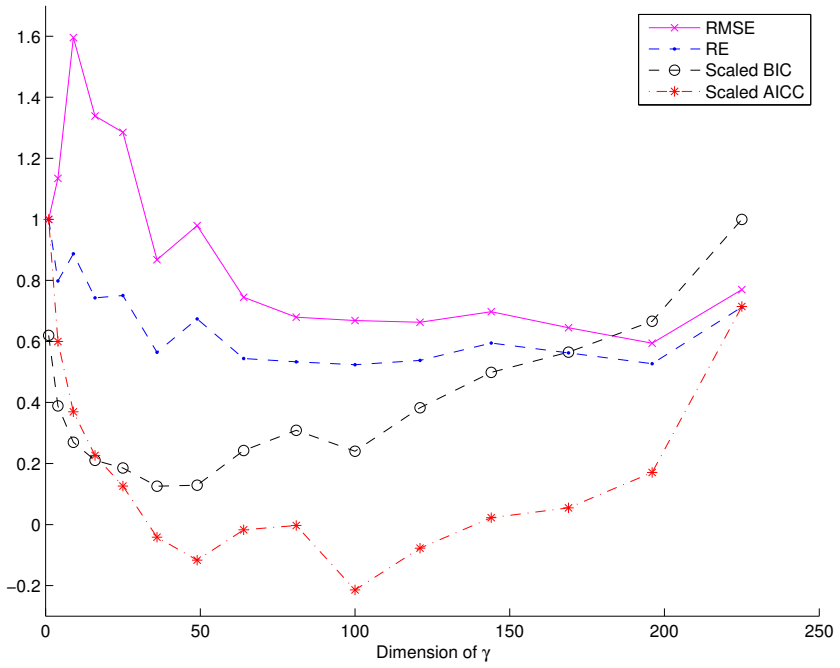
Compared to the QMLE and SMLE, our DS-SMLE estimator does not restrict the dependence structure but uses a sparsity scenario, that is, it estimates only non-zero elements of γ. For all three estimators, we report bias, variance, MSE, relative efficiency (RE) with respect to the QMLE and relative MSE (RMSE) with respect to QMLE. For SMLE and DS-SMLE we also report the dimension of γ. The number of observations is 500 and the number of replications is 1,000.

We consider four data generation processes. All have the same exponential marginals, where the mean μ is the parameter of interest with the true value $\mu_1 = \mu_2 = 0.5$, but the copula functions are different. We use the Plackett ,Student-t, Frank, and Gaussian copula as true copula. The copula parameter varies over the relevant range, representing different strengths of dependence. We report Kendall's τ for each such value.

We use Akaike information criterion (AIC) to choose the dimension of γ. The simulation results show that AIC is a more reasonable indicator than Bayesian

information criterion (BIC). Figure 2.2 illustrates that models with the low RE and RMSE usually have smaller AIC scores.

Figure 2.2: AIC, BIC, RE and RMSE for different values of $dim(\gamma)$



Notes: The DGP is exponential marginals and Frank Copula with Kendall's $\tau = -0.8$. The number of replications is 1000 and the number of observations is 500. BIC and AIC in the figure are the average of simulated values. The circle line is scaled BIC; the star dash line is scaled AIC; the dot dash line is relative efficiency; the solid line is relative mean squared errors.

Table 2.2 summarizes the simulation results. Two things are important here. First, for some values of τ, the DS-SMLE is at least as efficient as unrestricted SMLE, while it dramatically reduces the number of sieve parameters to be estimated. For example, Table 2.2 shows that ,when $\tau = -0.9$ with Plackett copula,the DS-SMLE estimates only 15 of 256 sieve parameters and it preserves the efficiency gains of the SMLE. We can observe simiar patterns if DGP is Frank, Student t, or Gaussian copula. Second, as negative dependence goes from high to low, both the SMLE and DS-SMLE have decreasing relative efficiency over QMLE. For instance, Table 2.2 shows as Kendall's τ varies from -0.9 to -0.7, both of DS-SMLE and SMLE become less advantageous than QMLE.

Table 2.1: QMLE, t copula based Pseudo-MLE, SMLE, DS-SMLE for insurance application with standard errors

	QMLE (Rob.St.Er)	PMLE (Rob.St.Er)	SMLE (St.Er)	DS-SMLE (St.Er)
a	14.7561	15.0103	15.7039	15.0344
	(4.4702)	(4.3306)	(3.1607)	(3.4653)
b	9.7020	9.6806	9.2871	9.7482
	(2.9080)	(2.8499)	(2.1433)	(2.4846)
LogL	-290.8190	-266.3389	-271.5390	-271.7004

2.3 Application from Insurance

We illustrate the use of the DS-SMLE with an insurance application. We consider automobile bodily injury liability claims from a sample of $n = 29$ Massachusetts towns in 1995 and 1997. The details of the data set can be found in Frees and Wang (2005). The two cross-sections have a strong positive correlation at 0.88 in the average town-wide claims (AC).

Following Frees and Wang (2005), the claims are assumed to have the same gamma distribution for the two years and the goal is the efficient estimation of the parameters (a, b) of that distribution. That is, we use the following cdf and pdf, respectively:

$$\begin{aligned} F_i(x|a,b) &= \frac{1}{b^a\Gamma(a)}\int_0^x t^{a-1}e^{\frac{-t}{b}}dt \\ f_i(x|a,b) &= \frac{1}{b^a\Gamma(a)}x^{a-1}e^{\frac{-x}{b}} \qquad \text{where i=1,2} \end{aligned}$$

The four estimators we consider are QMLE, Pseudo-MLE (PMLE), SMLE, and DS-SMLE. The QMLE estimator assumes independence between cross-section. It is known to be consistent even if the independence assumption is incorrect. To obtain a robust estimator of the standard errors, the "sandwich" formula is used. The PMLE is the estimator based on a fully specified parametric joint likelihood. We follow Frees and Wang (2005) and use t-copula for this. The PMLE is consistent

if the assume copula family is correct. Otherwise, the PMLE is generally biased and we do not know either the sign or the magnitude of the bias. Both the SMLE and DS-SMLE are robust in the sense that they do not depend on a specific assumption on the copula family. They are more efficient asymptotically relative to QMLE and, as illustrated by the simulation of the previous sections, behavie similarly in small samples.

Table 2.1 reports the estimates and standard errors. A few interesting observations can be made using these results. First, both the SMLE and DS-SMLE have smaller standard errors than QMLE. Second, while the SMLE shows evidence of bias, the DS-SMLE estimates are fairly close to FMLE or QMLE. We use 8 parameters in each dimension of the sieve, where this value is chosen using the BIC criterion. So for the SMLE, we have 66 parameter to estimate. For the DS-SMLE, we have only 10.

2.4 Concluding Remarks

We have proposed to use a penalized sieve to improve efficiency of likelihood-based estimators in panel settings. The settings can easily be generalized to multivariate models where a part of the joint distribution is modelled by a sieve with a potentially very large number of parameters, only a few of which are significant.

We show that the sparse sieve MLE, based on the Dantzig penalization, has very similar properties to the sieve MLE in finite samples, so the sparsity imposed by the Dantzig constraint does not add to the bias as much as much as it takes away from the variance. We also looked at the behavior of the estimator for various values of the tolerance and found evidence that our estimator tends to over-shrink. We proposes a two-step procedure that addresses this issue and clarifies the problem of choosing the tolerance level.

The relative efficiency and mean square gains we obtain are up to 70% which is very encouraging. The computational benefit is of course even more important; especially in cases when SMLE is infeasible due to small sample size.

Table 2.2: Simulation of SMLE and DS-SMLE for 3 dependence levels, using 1000 simulations with true copula to be Placket, Student's t, Frank and Gaussian copula. The sample size for each simulation is 500

		Plackett			Student's t			Frank			Gaussian		
Kendall's τ	Statistics	QMLE	SMLE	DS-SMLE	QMLE	SMLE	DS-SMLE	QMLE	SMLE	DS-SMLE	QMLE	SMLE	DS-SMLE
$\tau = -0.9$	mean	0.5005	0.4934	0.4954	0.5016	0.4941	0.4939	0.4996	0.4926	0.4927	0.5005	0.4925	0.4928
	NVar	0.5208	0.2169	0.1608	0.4983	0.1404	0.1561	0.4772	0.1479	0.1552	0.5100	0.1725	0.1781
	MSE	0.5206	0.2605	0.1816	0.5003	0.1747	0.1933	0.4769	0.2032	0.2077	0.5097	0.2290	0.2298
	$dim(\gamma)$		256	12		225	150		225	111		144	10
	RE		0.4164	0.3087		0.2817	0.3132		0.3099	0.3252		0.3382	0.3492
	RMSE		0.5005	0.3488		0.3491	0.3864		0.4262	0.4355		0.4492	0.4507
$\tau = -0.8$	mean	0.4999	0.4903	0.4941	0.499	0.4933	0.4927	0.4996	0.4879	0.4947	0.4986	0.4902	0.4900
	NVar	0.4836	0.3491	0.2888	0.4889	0.2904	0.243	0.4790	0.3131	0.2787	0.4732	0.3609	0.3017
	MSE	0.4831	0.4433	0.3228	0.4895	0.3356	0.2967	0.4787	0.4580	0.3062	0.4746	0.4571	0.4014
	$dim(\gamma)$		49	8		225	64		49	7		49	14
	RE		0.722	0.5972		0.594	0.4971		0.6538	0.5819		0.7627	0.6375
	RMSE		0.9177	0.6682		0.6857	0.6062		0.9569	0.6397		0.9632	0.8458
$\tau = -0.7$	mean	0.5006	0.4927	0.4935	0.4991	0.4945	0.4947	0.5006	0.4896	0.4894	0.5008	0.4949	0.4959
	NVar	0.5101	0.4456	0.4461	0.4953	0.4377	0.4017	0.5230	0.3633	0.3507	0.5096	0.4664	0.4309
	MSE	0.5099	0.4986	0.4875	0.4956	0.4671	0.429	0.5229	0.4714	0.4623	0.5097	0.4919	0.4470
	$dim(\gamma)$		36	6		64	62		36	13		64	9
	RE		0.8735	0.8747		0.8836	0.811		0.6946	0.6704		0.9154	0.8456
	RMSE		0.9779	0.9561		0.9425	0.8656		0.9015	0.8842		0.9651	0.8770

Note: SMLE, Sieve MLE; DS-SMLE, Dantzig Selector based Sieve MLE; RE, Relative Efficiency with respect to QMLE; RMSE, Relative MSE with respect to QMLE; $dim(\gamma)$, dimension of γ

Chapter 3

On Asymptotic Efficiency of Improved QMLE and Sieve MLE

Consider the setting of a panel with T time periods and N individuals. Assume T is fixed and $N \to \infty$. Suppose that for each cross section, we have a correctly specified parametric likelihood-based model and we can estimate this model consistently using only the cross sectional data. However, it is usually possible to use the entire panel to obtain more efficient estimators.

In this chapter we will focus on two such estimators. One estimator, known as the Improved QMLE (IQMLE), was proposed by Prokhorov and Schmidt (2009b). It is based on optimally weighted marginal score functions from all T cross sections. More specifically, let y_{it} represent the data for person i at time t, $t = 1, \ldots, T$, $i = 1, \ldots, N$. Assume that y_{it} are iid over i but not necessarily over t. Let $\ln f(y_{it}; \beta)$ denote the log-density of y_{it}, where β collects the unknown density parameters. This is the density we assume we have correctly specified. It is a marginal density in the sense that it corresponds to the marginal (single t) distribution of an unknown joint distribution over all t's. The derivative of the log-density with respect to β, known as the marginal score function, is a zero

mean function when evaluated at the true value of β. That is,

$$\mathbb{E}s_{it}(\beta_o) = 0, \text{ any } t,$$

where $s_{it}(\beta) = \nabla \ln f(y_{it};\beta)$ and ∇ denotes the derivative with respect to β. Prokhorov and Schmidt (2009b) propose to estimate β by optimal GMM based on all T zero-mean score functions. That is, their estimator is based on the following moment conditions

$$\mathbb{E}\begin{bmatrix} s_{i1}(\beta_o) \\ \dots \\ s_{iT}(\beta_o) \end{bmatrix} = 0 \tag{1}$$

The alternative estimator we consider is the sieve MLE (SMLE) proposed by Panchenko and Prokhorov (2013). In essence the SMLE is a maximum likelihood estimator which uses an approximator for the true joint log-density. Let $h(y_{i1,},\dots,y_{iT};\beta)$ denote the joint density of $(y_{i1,},\dots,y_{iT})$ and let $c(u_1,\dots,u_T)$ denote the copula density, corresponding to the joint density $h(y_{i1,},\dots,y_{iT};\beta)$ and the marginals $f(y_{it};\beta), t = 1,\dots,T$. It is a well known result (see, e.g, Sklar, 1959) that

$$\ln h(y_{i1,},\dots,y_{iT};\beta) = \sum_{t=1}^{T} \ln f(y_{it};\beta) + \ln c(F(y_{i1};\beta),\dots,F(y_{iT};\beta)), \tag{2}$$

where $F(y_{it};\beta)$ denotes the relevant cdf. The SMLE uses the sieve approximator of this joint log-density. That is, it replaces the last term in (2) with a truncated infinite series representation (a sieve) of the copula log density and then carries out the usual optimization over both β and the parameters of that representation. Denote the sieve parameters by γ_N and the sieve approximator by $\ln c_\gamma$. Then, the

SMLE maximizes the approximate joint log likelihood

$$\ln L_\gamma(\beta) = \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \ln f(y_{it};\beta) + \ln c_\gamma(F(y_{i1};\beta),\ldots,F(y_{iT};\beta)) \right].$$

Under appropriate regularity conditions, both estimators are consistent and asymptotically normal. Moreover, the IQMLE is known to be asymptotically efficient among the regular estimators that use moment conditions (1), while the SMLE is known to reach a semiparametric efficiency bound for estimation of β. However, their asymptotic variance matrices are not identical and it is unclear why one or the other should be preferred and under what circumstances.

The rest of this chapter is organized as follows. Section 3.1 introduces the asymptotic variance of IQMLE and SMLE. Section 3.2 presents the theoretical results regarding the relative efficiency of SMLE to IQMLE. Section 3.3 provides the simulation results. Finally, Section 3.4 concludes.

3.1 Asymptotic Variance of IQMLE and SMLE

This section discusses asymptotic properties of IQMLE, SMLE and related estimators. We assume without loss of generality that $T = 2$. This assumption is standard for copula literature and is used just to keep the notation under control.

3.1.1 Assumptions

Let $\theta \equiv (\beta', c)' \in \Theta \equiv B \times \Gamma$, where $B \subset \mathbb{R}^p$ and Γ is a space of copula functions. Let $\theta_N \equiv (\beta', \gamma_N)' \in \Theta_N \equiv B \times \Gamma_N$, where Γ_N is a sieve space for Γ. We need assumptions that ensure consistency and asymptotic normality of both SMLE and GMM based on the marginal scores.

Assumption 3.1. *(identification) $\beta_o \in int(B)$, B is compact and there exists a unique β_o for which equation (1) holds and that is the same β_o which maximizes $\mathbb{E}[\ln h(\mathbf{y}_i;\beta)]$ where $\mathbf{y}_i = (y_{i1}, y_{i2})$ and $\ln h$ is defined in (2).*

Assumption 3.2. *(smoothness)* $\Gamma = \{c = \exp(g) : g \in \Lambda^r([0,1]^m), r > 1/2, \int c(\mathbf{u})d\mathbf{u} = 1\}$, *where* $\mathbf{u} = (u_1, u_2)$, $\Lambda^r([0,1]^2)$ *denotes the class of* r*-smooth (Hölder) functions on* $[0,1]^2$, *and* $\ln f_j(y_j;\theta), j = 1,2$, *are twice continuously differentiable w.r.t.* θ.

Also, for SMLE asymptotics assume Assumptions 3-4 of Panchenko and Prokhorov (2013) and for the IQMLE asymptotics assume the equivalents of Assumptions of 2.1.iv-vi of Prokhorov and Schmidt (2009a).

3.1.2 QMLE and Improved QMLE

Recall the moment conditions

$$\mathbb{E}s_i^*(\beta_o) = 0, \quad \text{where} \quad s_i^*(\beta) = \begin{bmatrix} s_{i1}(\beta) \\ s_{i2}(\beta) \end{bmatrix}. \tag{3}$$

The IQMLE is an optimal GMM estimator based on the moment conditions in (3).[1] It should be consistent so long as the marginal distributions are correctly specified, since $\mathbb{E}s_i^*(\beta_o) = 0$ if $\mathbb{E}s_{it}(\beta_o) = 0$ for all t.

Define $\mathbb{H}_* = \mathbb{E}\nabla_\beta s_i^*(\beta_o)$ and $\mathbb{V}_* = \mathbb{E}s_i^*(\beta_o)s_i^*(\beta_o)'$. Then, if $\hat{\beta}_{\text{IQMLE}}$ is the IQMLE estimator, standard results would indicate that the asymptotic variance of $\hat{\beta}_{\text{IQMLE}}$ is

$$\mathbb{V}_{\text{IQMLE}} = (\mathbb{H}_*'\mathbb{V}_*^{-1}\mathbb{H}_*)^{-1}.$$

It is a well known result that $\hat{\beta}_{\text{IQMLE}}$ is asymptotically efficient in the class of estimators using the moment conditions (3), that is, no other regular estimator using the same moment conditions achieves an asymptotical variance smaller than $\mathbb{V}_{\text{IQMLE}}$ (see, e.g., Newey and McFadden, 1994). For example, the traditional

[1] Because we deal only in asymptotics here, we will not be explicit about issues like how to estimate the optimal GMM weighting matrix. Nor would it matter if instead of GMM we considered other asymptotically equivalent estimates, such as empirical likelihood, exponential tilting, etc.

QMLE [2] based on the moment condition

$$\mathbb{E}[s_{i1}(\beta_o) + s_{i2}(\beta_o)] = 0,$$

is dominated by IQMLE because optimal GMM weighting dominates summation.

Prokhorov and Schmidt (2009b, Section 2) derive conditions under which IQMLE and QMLE are equally efficient. They also prove that the matrices $\mathbb{H}_*$ and $\mathbb{V}_*$ have the following structure:

$$\mathbb{V}_* = \begin{bmatrix} A & G \\ G' & B \end{bmatrix}, \mathbb{H}_* = \begin{bmatrix} -A \\ -B \end{bmatrix} \tag{4}$$

where $A = \mathbb{E}s_{i1}(\beta_o)s_{i1}(\beta_o)'$, $B = \mathbb{E}s_{i2}(\beta_o)s_{i2}(\beta_o)'$ and $G = \mathbb{E}s_{i1}(\beta_o)s_{i2}(\beta_o)'$.

3.1.3 Full MLE and Sieve MLE

We follow Panchenko and Prokhorov (2013) and use the Bernstein polynomial sieve introduced by Sancetta and Satchell (2004):

$$c_\gamma(\mathbf{u}) = J_N^2 \sum_{v_1=0}^{J_N-1} \sum_{v_2=0}^{J_N-1} \omega_v \prod_{l=1}^{2} \begin{pmatrix} J_N - 1 \\ v_l \end{pmatrix} u_l^{v_l}(1-u_l)^{J_N-v_l-1}, u \tag{5}$$

where $\gamma = \{\omega_v\}$ denotes parameters of the polynomial indexed by $v = (v_1, \ldots, v_m)$ such that $0 \leq \omega_v \leq 1$ and $\sum_{v_1=0}^{J_N-1} \sum_{v_2=0}^{J_N-1} \omega_v = 1$. Sancetta (2007) derives rates of convergence of the Bernstein copula to the true copula. Ghosal (2001) and references therein discuss the rate of convergence of the sieve MLE based on the Bernstein polynomial (only for one-dimensional densities). Other sieves can be used (see, e.g., Bouezmarni and Rombouts, 2010; Chen, 2007) but we found this sieve to perform better in simulations than others.

Note 3.1. *For the Bernstein copula sieve, Γ_N is dense in Γ, i.e. $c_\gamma \to c_o$ as*

[2] QMLE represent Quasi Maximum Likelihood Estimation, which constructs the joint likelihood assuming independence. To distinguish, the QMLE is different from Quasi Likelihood which is an estimating function(see, e.g., Heyde, 1997).

$J_N \to \infty$; the dimension of γ is J_N^2 and the dimension of θ_N is $p + J_N^2$.

The sieve MLE can be written as

$$\hat{\theta}_{\text{SMLE}} = \arg \max_{\theta_N \in \Theta_N} \sum_{i=1}^{N} \left[\ln f(y_{i1}; \beta) + \ln f(y_{i2}; \beta) + \ln c_\gamma(F(y_{i1}; \beta), F(y_{i2}; \beta)) \right], \quad (6)$$

where $c_\gamma(\cdot)$ is given in (5).

If there exists a value of γ_N, for which $c(u_1, u_2) = c_\gamma(u_1, u_2)$ for any $(u_1, u_2) \in [0, 1]^2$ then we have the true joint log-density in (6) and the SMLE becomes the full MLE, where γ_N are nuisance parameters. In this case, it is a well known result that the variance of $\hat{\beta}_{\text{SMLE}}$ – the variance of the first p elements of $\hat{\theta}_{\text{SMLE}}$ – can be obtained as the variance of the population error projection of the score with respect β on the space spanned by the score with respect to γ_N.

In general, c_γ is not equal to the true copula density for any (finite dimensional) γ_N due to the approximation error and the asymptotic variance derivation is harder (see, e.g., Ai and Chen, 2003; Chen, Fan, and Tsyrennikov, 2006; Chen and Pouzo, 2009; Panchenko and Prokhorov, 2013; Shen, 1997, who study the asymptotic properties of this estimator in various settings). We will now provide just the end results of this derivation; more details can be found in Panchenko and Prokhorov (2013).

For each component β_q, $q = 1, \ldots, p$, denote by g_q^* the solution to

$$\inf_{g_q} E \left[\sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_j, \beta_o)}{\partial \beta_q} + \left(\frac{1}{c(\mathbf{u})} \frac{\partial c(\mathbf{u})}{\partial u_j} \right) \Big|_{u_k = F_k(y_k, \beta_o)} \frac{\partial F_j(y_j, \beta_o)}{\partial \beta_q} \right\} \right. \quad (7)$$
$$\left. + \left(\frac{1}{c(\mathbf{u})} g_q(u_1, \ldots, u_m) \right) \Big|_{u_k = F_k(y_k, \beta_o)} \right]^2.$$

and define

$$S'_\beta = \sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_j, \beta_o)}{\partial \beta'} + \left(\frac{1}{c(\mathbf{u})} \frac{\partial c(u_1, \ldots, u_m)}{\partial u_j} \right) \Big|_{u_k = F_k(y_k, \beta_o)} \frac{\partial F_j(y_j, \beta_o)}{\partial \beta'} \right\} \quad (8)$$
$$+ \left(\frac{1}{c(\mathbf{u})} g^*(u_1, \ldots, u_m) \right) \Big|_{u_k = F_k(y_k, \beta_o)},$$

where $g^* = (g_1^*, \ldots, g_p^*)$. Then, Panchenko and Prokhorov (2013) show that under the assumptions above, $\sqrt{N}(\hat{\beta}_{\text{SMLE}} - \beta_o) \to N(0, \mathbb{V}_{\text{SMLE}})$, where

$$\mathbb{V}_{\text{SMLE}} = (\mathbb{E}[S_\beta S_\beta'])^{-1},$$

and $\hat{\beta}_{\text{SMLE}}$ is semiparametrically efficient.

The efficiency result suggests that $\hat{\beta}_{\text{SMLE}}$ has the smallest asymptotic variance, in the positive definite sense, among all regular estimators which use the information contained in the parametric marginals. The SMLE efficiency argument goes back to Stein (1956). He observed that a model with a nonparametric component is at least as hard as any one-dimensional submodel that satisfies semiparametric assumptions. If we "parameterize" our problem as $\theta(t) = \theta_o + t\nu$, where $\nu \in \bar{V}(\Theta, \theta_o)$ is the closed linear span of $\Theta - \{\theta_o\}$ also known as the tangent space, then estimating θ_o should be at least as hard as estimating t ("true" $t = 0$). The semiparametric lower bound, reached by SMLE, is then the supremum over the traditional Cramer-Rao bounds (infimum over Fisher informations) for estimating t over all suitable parametric submodels. Therefore, $\mathbb{V}_{\text{SMLE}}$ is no smaller, in positive definite sense, than the asymptotic variance matrix obtained using a full parametric MLE. At the same time, $\mathbb{V}_{\text{SMLE}}$ is not generally the same as (and should be smaller than) $\mathbb{V}_{\text{IQMLE}}$ and, consequently, $\mathbb{V}_{\text{QMLE}}$.

Note 3.2. *If $J_n = 1$, SMLE reduces to QMLE. In this case, the sieve copula pdf contains only a constant and can be disregarded in estimation, which gives the QMLE objective function.*

We will now define new notation that will make the Stein (1956) argument formal. Define the directional derivative of the log-likelihood in direction $\nu =$

$(\nu'_\beta, \nu_\gamma)' \in V$, where V is the linear span of $\Theta - \{\theta_o\}$,

$$\begin{aligned}
\dot{l}(\theta_o)[\nu] &\equiv \lim_{t\to 0} \frac{\ln h(y,\theta+t\nu) - \ln h(y,\theta)}{t}\Big|_{\theta=\theta_o} \\
&= \frac{\partial \ln h(y,\theta_o)}{\partial \theta'}[\nu] \\
&= \sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_j,\beta_o)}{\partial \beta'} + \left(\frac{1}{c(u_1,\ldots,u_m)} \frac{\partial c(u_1,\ldots,u_m)}{\partial u_j} \right)\Big|_{u_k=F_k(y_k,\beta_o)} \frac{\partial F_j(y_j,\beta_o)}{\partial \beta'} \right\} \nu_\beta \\
&\quad + \frac{1}{c(F_1(y_1,\beta_o),\ldots,F_m(y_m,\beta_o))} \nu_\gamma(u_1,\ldots,u_m)\Big|_{u_k=F_k(y_k,\beta_o)}
\end{aligned}$$

Also, define the Fisher inner product $\langle \cdot, \cdot \rangle \equiv \mathbb{E}\left[\dot{l}(\theta_o)[\cdot]\dot{l}(\theta_o)[\cdot]\right]$ on space V and the Fisher norm $||\nu|| \equiv \sqrt{\langle \nu, \nu \rangle}$, where expectation is with respect to the true density h. The closed linear span of $\Theta - \{\theta_o\}$ and the inner product $\langle \cdot, \cdot \rangle$ form a Hilbert space, call it $(\bar{V}, ||\cdot||)$.

Note 3.3. *The Fisher information for estimating t is now the Fisher norm $\mathbb{E}\left[\dot{l}(\theta_o)[\cdot]\dot{l}(\theta_o)[\cdot]\right]$.*

Given the consistent estimates $\hat{\beta}_{\text{SMLE}}$ and $\hat{c}_{\text{SMLE}} = c_{\hat{\gamma}_N}$, g_q^*'s can be estimated consistently in a sieve minimization problem as follows

$$\arg\min_{g_q \in \mathbf{A}_N} \sum_{i=1}^{N} \left[\sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_{ji},\hat{\beta})}{\partial \beta_q} + \left(\frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} \frac{\partial \hat{c}(\hat{u}_{1i},\ldots,\hat{u}_{mi})}{\partial u_j} \right)\Big|_{\hat{u}_{ki}=F_k(y_{ki},\hat{\beta})} \frac{\partial F_j(y_{ji},\hat{\beta})}{\partial \beta_q} \right\} + \frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} g_q(\hat{u}_{1i},\ldots,\hat{u}_{mi})\Big|_{\hat{u}_{ki}=F_k(y_{ki},\hat{\beta})} \right]^2,$$

where $q = 1,\ldots,p$ and $\mathbf{A}_N$ is one of the sieve spaces discussed above. Given consistent estimates $\hat{\beta}$, $\hat{c}$, and $\hat{g}^*$, a consistent estimate of $E[S_\beta S'_\beta]$ is easy to obtain if we replace the expectation evaluated at the true values with a sample average evaluated at the estimates.

A simpler alternative estimator of the sieve MLE asymptotic variance is provided by Ackerberg, Chen, and Hahn (2009). They show that one can use the upper left $p \times p$ block of the usual MLE covariance matrix as an estimate of $(E[S_\beta S'_\beta])^{-1}$ provided that the outer-product-of-the-score form of the covariance matrix is used.

Note 3.4. *The variance estimator of Ackerberg, Chen, and Hahn (2009) can be viewed as a Pseudo-MLE (PMLE) variance estimator.*

3.2 IQMLE versus SMLE

Before comparing IQMLE with SMLE directly, we will first consider the cases where IQMLE is equally efficient as QMLE. Once we can identify these cases, comparing SMLE with IQMLE is reduced to compare SMLE with QMLE.

Then we will compare IQMLE with SMLE when the number of sieve parameters is fixed. In other words, suppose we use the Bernstein copula with J_n fixed at a certain number k. Then the SMLE becomes a fully parametric problem with $p+k^2$ parameters to estimate. The joint density used in this problem is not the correct density h, but a pseudo-density, which would be h if J_n was infinity. This kind of comparison is useful for practitioners, who regard SMLE as a PMLE, whose variance matrix is estimated without using the "sandwich" formula to account for the joint density misspecification.

Finally, we will compare IQMLE with SMLE. We consider a simple example of a bivariate discrete model and show that SMLE dominates IQMLE in this specific setting.

3.2.1 IQMLE vs QMLE

We consider three situations: same marginals with common parameters, same marginals with distinct parameters, and different marginals with distinct parameters. Theorem 3.1 compares IQMLE and QMLE when marginals do not share common parameters, while Theorem 3.2 compares them in a contrary case.

Theorem 3.1. *If the marginal pdf have distinct parameters, IQMLE is the same as QMLE.*

Proof. see Appendix 3.5 for all proofs. □

Theorem 3.2. *(Prokhorov and Schmidt, 2009b) If the marginal pdf have common parameters, IQMLE estimator is efficient relative to QMLE estimator, and the two estimators are equally efficient if and only if H_* is in the space spanned by (V_*A')*

3.2.2 IQMLE vs SMLE fixed k

3.2.2.1 Asymptotic variance for SMLE fixed k

In this section, we no longer impose any restrictions on marginal pdf parameters. We view SMLE as a parametric MLE except that the sieve copula is generally not the true copula family. More precisely, we estimate $\hat{\theta}_N = (\hat{\beta}_{SMLE-K}, \hat{\gamma}_N)' = \arg\max_{\theta \in \Theta_N} E[l(z_i, \beta, \gamma)]$, where Θ_N is the sieve of Θ.

$$\sqrt{n}(\hat{\beta}_{SMLE} - \beta_*, \hat{\gamma} - \gamma_*)' \to N(0, I^{-1}) \tag{9}$$

$$\text{where, } I = \begin{bmatrix} E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\beta} \frac{dl_s(z,\beta_*,\gamma_*)}{d\beta'}] & E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\beta} \frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma'}] \\ E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma} \frac{dl_s(z,\beta_*,\gamma_*)}{d\beta'}] & E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma} \frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma'}] \end{bmatrix} \tag{10}$$

$$l_s = \ln f_1 + \ln f_2 + \ln c_\gamma \tag{11}$$

The asymptotic variance of $\hat{\beta}_{SMLE-K}$ is the upper left block in the inverse of the covariance matrix. One remark is that the generalized information matrix equality does not apply to I because the sieve copula $\ln c_\gamma$ is not the true copula $\ln c$.

Equation (4) shows that the asymptotic variance of IQMLE depends on the covariance matrix of moment conditions in (3). Accordingly, we would like to represent the asymptotic variance of SMLE(fixed k) in a similar way as IQMLE, except that the moment conditions for SMLE(fixed k) are different.

We have four moment conditions based on sieve MLE (fixed k) estimation. In addition, if the sieve copula is the same as the true copula, we have a correctly specified full MLE model and another group of moment conditions. Accordingly, we can define the covariance matrix in an explicit way.

$$\begin{array}{rcl} E(\nabla_\beta ln f_1) = 0 & & E(\nabla_\beta ln f_1) = 0 \\ E(\nabla_\beta ln f_2) = 0 & \underset{\Longrightarrow}{c_s=c} & E(\nabla_\beta ln f_2) = 0 \\ E(\nabla_\beta ln c_s) = 0 & & E(\nabla_\beta ln c) = 0 \\ E(\nabla_\gamma ln c_s) = 0 & & E(\nabla_\gamma ln c) = 0 \end{array} \tag{12}$$

Definition 3.1. *Let C_{sieve} be the variance covariance matrix for moments (12) of*

sieve MLE with fixed k, and C_{true} be the covariance matrix for moments when we have correct full MLE model. Then C_{sieve} and C_{true} can be expressed as

$$C_{sieve} = \left[\begin{array}{ccc|c} A & G & -K & -P \\ G' & B & -L' & -Q' \\ -K' & -L & N & Z \\ \hline -P' & -Q & Z' & W \end{array}\right] \quad C_{true} = \left[\begin{array}{ccc|c} A & G & -G & 0 \\ G' & B & -G' & 0 \\ -G' & -G & J & E \\ \hline 0 & 0 & E' & F \end{array}\right] \tag{13}$$

where, A,B,G,K,L N,P,Q ,Z,W,J,E,F are defined in Definition 3.2 in Appendix.

We first decompose $E(\nabla_\beta l_s \nabla'_\beta l_s)$,$E(\nabla_\beta l_s \nabla'_\gamma l_s)$, and $E(\nabla_\theta l_s \nabla'_\gamma l_s)$ in terms of covariance of moments in Definition 3.1, and Lemma 3.1 derives another representation of information matrix I. Then Lemma 3.2 uses the fact that the asymptotic variance of SMLE fixed k is just the upper-left block in the inverse of the information matrix.

Lemma 3.1. *The information matrix I in SMLE fixed k can be written as*

$$I = \begin{bmatrix} A+G+G'+B-K-L'-K'-L+N & -P-Q'+Z \\ -P'-Q+Z' & W \end{bmatrix} \tag{14}$$

Lemma 3.2. *The asymptotic variance of SMLE(fixed k) is the upper left block in the inverse of information matrix I.*

$$\begin{aligned} \mathbb{V}_{\text{SMLE-K}} &= [A+G+G'+B-K-L'-K'-L+N \\ &\quad -(-P-Q'+Z)W^{-1}(-P'-Q+Z')]^{-1} \end{aligned} \tag{15}$$

3.2.2.2 Theoretical results and intuition

First, we give the expression of the asymptotic variance of IQMLE.

Lemma 3.3. *(Prokhorov and Schmidt, 2009b) The asymptotic variance of IQMLE is*

$$\mathbb{V}_{\text{IQMLE}} = \{\begin{bmatrix}-A & -B\end{bmatrix}\begin{bmatrix}A & G\\ G' & B\end{bmatrix}^{-1}\begin{bmatrix}-A\\ -B\end{bmatrix}\}^{-1} \tag{16}$$

$$= \{A+B-G'-G+\begin{bmatrix}-G' & -G\end{bmatrix}\begin{bmatrix}A & G\\ G' & B\end{bmatrix}^{-1}\begin{bmatrix}-G\\ -G'\end{bmatrix}\}^{-1} \tag{17}$$

Theorem 3.3. *IQMLE is asymptotically as efficient as SMLE with fixed number of sieve parameters if and only if the following condition is satisfied:*

$$A+G+G'+B-K-L'-K'-L+N-(-P-Q'+Z)W^{-1}(-P'-Q+Z') = A+B-G'-G+\begin{bmatrix}-G' & -G\end{bmatrix}\begin{bmatrix}A & G\\ G' & B\end{bmatrix}^{-1}\begin{bmatrix}-G\\ -G'\end{bmatrix} \tag{18}$$

Given the expression of asymptotic variance of IQMLE and SMLE(fixed k), Theorem 3.3 derives a sufficient and necessary condition for when the two estimators are equally asymptotically efficient. The equal efficiency condition in Theorem 3.3 is difficult to interpret. In fact, we can represent it in an intuitive way. We proceed in the following order. First, in Lemma 3.4 and 3.5 we represent $\mathbb{V}_{\text{IQMLE}}$ and $\mathbb{V}_{\text{SMLE-K}}$ in a intuitive way such that they are general enough to include different values of T. Then, in Theorem 3.4 we establish a necessary and sufficient condition under which SMLE(fixed k) is equally efficient as IQMLE. Finally, in Theorem 3.5 we make a link to a previous result in Prokhorov and Schmidt (2009b).

Lemma 3.4. *The asymptotic variance for IQMLE can be expressed as*

$$\mathbb{V}_{\text{IQMLE}} = \{E(\nabla_\beta l\nabla'_\beta l) - E(\nabla_\beta lnc\nabla'_\beta lnc) + E(\widetilde{\nabla_\beta lnc}\widetilde{\nabla'_\beta lnc})\}^{-1} \tag{19}$$

$$= \{E(\nabla_\beta l\nabla'_\beta l) - E(ee')\}^{-1} \tag{20}$$

where $l = lnf_1 + lnf_2 + lnc$, so l is the true log-density, e is the regression error of $\nabla_\beta \ln c$ regressed on $\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}$, $\widetilde{\nabla_\beta lnc}$ is the fitted value of regression of

$\nabla_\beta lnc$ *on* $\left[\nabla'_\beta lnf_1 \quad \nabla'_\beta lnf_2\right]$.

There are several interesting remarks on $\mathbb{V}_{\text{IQMLE}}$ in Lemma 3.4.

First, although the proof of Lemma 3.4 is based on $T = 2$, it can be easily extended to include the cases where T is larger than 2. For example, assume $T = 3$, $\mathbb{V}_{\text{IQMLE}} = \{E(\nabla_\beta l \nabla'_\beta l) - E(ee')\}^{-1}$, where $l = \sum_{i=1}^{3} lnf_i + lnc$ and e is the regression error of $\nabla_\beta \ln c$ regressed on $\left[\nabla'_\beta lnf_1 \quad \nabla'_\beta lnf_2 \quad \nabla'_\beta lnf_3\right]$.

Second, by Equation (19), $\mathbb{V}_{\text{IQMLE}}$ is negatively related to the variance of $\widetilde{\nabla_\beta lnc}$, the fitted values of copula scores regressed on the marginal scores. Although IQMLE only uses the stacked marginal scores as moment conditions, it does include some information on copula scores, which can be expressed as a linear combination of marginal scores. In other words, the linear relationship between $\nabla_\beta \ln c$ and the marginal scores brings efficiency gains to IQMLE.

Third, let's consider the case where copula scores is a perfect linear combination of marginal scores, ie, $E(ee') = 0$. By Theorem 4 in Prokhorov and Schmidt (2009b) or Theorem 3.5 later in this section , IQMLE is equally efficient as FMLE. Since FMLE dominate SMLE and QMLE, IQMLE should also dominate SMLE and QMLE in this special case.

Lemma 3.5. *The asymptotic variance of SMLE(fixed k) can be expressed as:*

$$\begin{aligned} \mathbb{V}_{\text{SMLE-K}} &= \{E(\nabla_\beta l_s \nabla'_\beta l_s) - E(\nabla_\beta l_s \nabla'_{\gamma_N} l_s) E(\nabla_{\gamma_N} l_s \nabla'_{\gamma_N} l_s)^{-1} E(\nabla_{\gamma_N} l_s \nabla'_\beta l_s)\}^{-1} \\ &= \{E(\nabla_\beta l_s \nabla'_\beta l_s) - E(\widehat{\nabla_\beta l_s} \widehat{\nabla'_\beta l_s})\}^{-1} \end{aligned} \tag{21}$$

where $l_s = \ln f_1 + \ln f_2 + \ln c_s$ *and* $\widehat{\nabla_\beta l_s}$ *is the fitted value of regression of* $\nabla_\beta l_s$ *on* $\nabla'_{\gamma_N} l_s$.

We have some interesting observations on $\mathbb{V}_{\text{SMLE-K}}$ in Lemma 3.5.

First, Equation (21) shows that $\mathbb{V}_{\text{SMLE-K}}$ depends on how well the sieve likelihood l_s could approximate the true likelihood. However, the goodness of fit for the sieve copula is a two-edged sword. On the one hand, good approximation of the true likelihood will increase the variance of $\nabla_\beta l_s$, which will increase the efficiency

of SMLE. On the other hand, to increase the goodness of fit for l_s, we have to increase the number of sieve parameters γ_N. As a result, $\nabla_\beta l_s$ will be better fitted as we increase the number of regressor γ_N, which in turn implies an increase of $E(\widehat{\nabla_\beta l_s}\widehat{\nabla'_\beta l_s})$ and a decrease the efficiency of SMLE.

Second, SMLE has incorporated the sieve copula scores, which may not be a linear combination of the marginal scores. More precisely, compared with Equation (20) for IQMLE, Equation (21) shows that $\mathbb{V}_{\text{SMLE-K}}$ includes the variance of $\nabla_\beta l_s$ without subtracting $E(e_s e'_s)$, where e_s is the regression error of $\nabla_\beta \ln c_s$ regressed on $\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}$. In other words, $\mathbb{V}_{\text{SMLE-K}}$ encloses the part of $\nabla_\beta \ln c_s$ that could not be explained as a linear combination of marginal scores. Furthermore, if the sieve copula c_s can approximate the true copula well enough, SMLE could be viewed as containing some partial information on true copula scores that are not explained by the linear combination of marginal scores.

Finally, assume l_s is the true likelihood, then Equation (21) gives the asymptotic variance of Full-MLE. That is to say, $\mathbb{V}_{\text{MLE}} = \{E(\nabla_\beta l \nabla'_\beta l) - E(\widehat{\nabla_\beta l}\widehat{\nabla'_\beta l})\}^{-1}$, where where $l = \ln f_1 + \ln f_2 + \ln c$ and $\widehat{\nabla_\beta l}$ is the fitted value of regression of $\nabla_\beta l$ on $\nabla'_\gamma l$.

Theorem 3.4. *IQMLE is asymptotically as efficient as SMLE with fixed number of sieve parameters if and only if the following condition is satisfied:*

$$E(\nabla_\beta l \nabla'_\beta l) - E(ee') = E(\nabla_\beta l_s \nabla'_\beta l_s) - E(\widehat{\nabla_\beta l_s}\widehat{\nabla'_\beta l_s}) \tag{22}$$

where $l = lnf_1 + lnf_2 + lnc$, $l_s = lnf_1 + lnf_2 + lnc_s$, e *is the regression error of* $\nabla_\beta lnc$ *regressed on* $\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}$, $\widehat{\nabla_\beta l_s}$ *is the fitted value of regression of* $\nabla_\beta l_s$ *on* $\nabla'_\gamma l_s$

There are several implication of Theorem 3.4.

If the true copula score is a linear combination of the marginal scores, IQMLE should be efficient relative to SMLE with fixed k. By Theorem 4 in Prokhorov and Schmidt (2009b), IQMLE dominate FMLE under this condition, which in turn implies IQMLE also dominate SMLE fixed k.

In contrast, suppose there is nonlinear relationship between copula scores and marginal scores , and the sieve copula approximate the true copula well enough such that $E(\nabla_\beta l \nabla'_\beta l) \approx E(\nabla_\beta l_s \nabla'_\beta l_s)$, SMLE fixed k could be efficient relative to IQMLE. More precisely, if $E(\nabla_\beta l \nabla'_\beta l) \approx E(\nabla_\beta l_s \nabla'_\beta l_s)$ and $E(\widehat{\nabla_\beta l_s}\widehat{\nabla'_\beta l_s})$ is smaller than $E(ee')$, in the negative definite sense, by Equation (22) SMLE fixed k is more efficient than IQMLE.

Actually, we have some simulation results to support these implications.

Prokhorov and Schmidt (2009b) have a result regarding equal relative asymptotic efficiency between IQMLE and FMLE. It can be shown that their result is a special case of Theorem 3.4. If we assume the sieve log-density l_s is the same as the true log-density l and apply Theorem 3.4, we get the same condition as in Prokhorov and Schmidt (2009b) for FMLE to be equally efficient as IQMLE.

Theorem 3.5. *IQMLE is asymptotically as efficient as Full-MLE iff the following condition is satisfied.*

$$E(\nabla_\beta lnc \nabla'_\beta lnc) = E(\nabla_\beta lnc^* \nabla'_\beta lnc^*) \tag{23}$$

where $\nabla_\beta lnc^*$ *is the fitted value of regression of* $\nabla_\beta lnc$ *on* $\begin{bmatrix} \nabla'_\beta lnf_1 & \nabla'_\beta lnf_2 & \nabla\gamma' lnc \end{bmatrix}$. *In other words,* $\nabla_\beta lnc$ *is a linear combination of* $\nabla_\beta lnf_1$, $\nabla_\beta lnf_2$ *and* $\nabla\gamma lnc$.

3.2.3 IQMLE versus SMLE in an Example

We compare IQMLE with SMLE in an bivariate example.

Example 3.1. *We have a bivariate discrete distribution. The marginal probability mass functions are Bernoulli. The probability mass function is summarized in Table 3.1*

Table 3.1: Bivariate Bernoulli Probability Mass Function

	y_1	y_0	
x_1	γ	$1-\mu-\gamma$	$1-\mu$
x_0	$1-\mu-\gamma$	$2\mu-1+\gamma$	μ
	$1-\mu$	μ	1

We want to estimate the marginal parameter μ by SMLE and IQMLE. Please note that SMLE is the same as FMLE in this setup.

Theorem 3.6. *In Example 3.1, SMLE is always asymptotically efficient relative to IQMLE.*

3.3 Simulation Results

We conduct four different kinds of simulations to support the theoretical results derived in Section 3.2.

First, in Section 3.3.1 we have a data generating process (DGP) with the same exponential marginals, but in SMLE and IQMLE implementation we do not impose the restrictions that marginal pdf parameters are the same. In other words, we assume distinct marginal parameters. By Theorem 3.1, IQMLE should be the same as QMLE in this situation. The simulation results do show that IQMLE is almost the same as QMLE.

Second, in Section 3.3.2 we have normal marginals and the Gaussian copula as the true copula. In this DGP, the copula scores are a linear combination of the marginal scores. By Theorem 3.4, IQMLE should be more efficient than SMLE. In fact, IQMLE should be the same as FMLE in this case. The simulation result do show that SMLE could not improve on IQMLE.

Third, in Section 3.3.3 the DGP is very similar to that in Section 3.3.2 except the true copula is replaced by the Frank copula. As a result, the copula scores are no longer the linear combination of the marginal scores. By Theorem 3.4, SMLE

should dominate IQMLE. The simulation results show that SMLE has improved the efficiency of IQMLE by about 50%.

Finally, in Section 3.3.4 we compare SMLE and IQMLE with some incorrectly specified copula-based likelihood models. The marginals are exponential distributions, while the dependence is modelled by the Plackett, Frank and Gaussian copulas. The simulation results show that SMLE is more efficient and robust compared to IQMLE and a group of misspecified copula models.

3.3.1 Same marginals, Distinct parameters

We consider a bivariate DGP with exponential marginals in which both the mean parameters μ_1 and μ_2 are set to 0.5. The dependence is modelled by the Frank copula, with $\gamma_0 = -39$. So we have negative dependence. The number of observations is 1000.

Table 3.2 contains simulation results. MSE is minimized at $J_n = 16$. So we are estimating 256 nuisance parameters in the sieve copula and 2 parameters of the marginals. To reduce estimation time, we use parallel computing toolbox of Matlab in Cirrus[3].

We use 1000 simulation runs and report the simulated mean of QMLE, IQMLE, SMLE, and Full MLE, their simulated variance and mean square error (MSE), and the simulated relative efficiency (RE) and relative MSE (RMSE). As in Joe (2005), RE is the ratio of the SMLE simulated variance to that of QMLE. RMSE is the ratio of the SMLE simulated MSE to that of QMLE.

The simulation results in Table 3.2 suggest that in this specific data setting, IQMLE and QMLE are almost the same, and SMLE has improved efficiency over IQMLE by about 60%. This result supports Theorem 3.1 and 3.4. First, since we do not assume the same marginal parameters, IQMLE is expected to be the same as QMLE asymptotically. Second, the copula scores in this DGP is not a

[3]Cirrus is a cluster of high performance computers at department of engineering in Concordia university.

linear combination of the marginal scores, by Theorem 3.4, we expect SMLE to be efficient relative to IQMLE.

We note that in the following simulations, we always impose the same marginal parameters in the implementation of IQMLE, SMLE, FMLE and QMLE.

3.3.2 Normal Marginals, Gaussian Copula

We consider a bivariate DGP with normal marginals in which $\mu = 5$ and variance $\sigma^2 = 1$. The dependence is modelled by the Gaussian Copula with parameter $\gamma = -0.9$. So there is negative dependence between the two cross-sections.

We estimate the common mean μ with $\sigma^2 = 1$ by QMLE, IQMLE, SMLE, and full MLE respectively. We impose the restriction that the two marginals are the same in the implementation of these four estimators.

The MSE of SMLE is minimized when $J_n = 1$. The simulation results in Table 3.3 suggest that IQMLE is significantly better than SMLE, and IQMLE is almost the same as FMLE. In addition, we observe that SMLE is the same as QMLE if $J_n = 1$.

Actually, this simulation result is as expected. Assume Normal marginal densities with $\sigma_1^2 = \sigma_2^2 = 1$ and $\mu_1 = \mu_2 = \mu$, and let the true copula be Gaussian with parameter γ. We have

$$\begin{aligned} f_1(y_1;\mu) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_1-\mu)^2}{2}} \\ f_2(y_2;\mu) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_2-\mu)^2}{2}} \\ c(F_1(y_1;\mu), F_2(y_2;\mu);\gamma) &= \frac{1}{\sqrt{1-\gamma^2}} e^{-\frac{\gamma(\gamma(y_1-\mu)^2+\gamma(y_2-\mu)^2-2(y_1-\mu)(y_2-\mu))}{2(1-\gamma^2)}} \end{aligned}$$

The relevant moment conditions for full MLE are

$$E\{Y_1 - \mu\} = 0 \tag{24a}$$

$$E\{Y_1 - \mu\} = 0 \tag{24b}$$

$$E\{-\frac{((Y_1 - \mu) + (Y_2 - \mu))\gamma}{\gamma + 1}\} = 0 \tag{24c}$$

$$E\{-\frac{\gamma(Y_1^2 + Y_2^2) + \mu(1-\gamma)^2(y_1 + y_2) - (1+\gamma^2)Y_1Y_2 + \gamma(\gamma^2 - 1) - \mu^2(1-\gamma)^2}{(\gamma - 1)^2(\gamma + 1)^2}\} = 0 \tag{24d}$$

In this setting, the copula scores are a linear combination of the marginal scores. We observe that (24c) is a linear combination of (24a) and (24b), so (24c) is also a linear combination of (24a) , (24b) and (24d). By Prokhorov and Schmidt (2009b), we know that IQMLE is as efficient as Full MLE. As SMLE could not be more efficient than Full MLE, SMLE should be less efficient than IQMLE.

3.3.3 Normal Marginals, Frank Copula

The DGP is the same as that in Section 3.3.2 except that the dependence is modelled by the Frank copula with $\gamma = -39$. The MSE of SMLE is minimized at $J_n = 5$. The simulation results in Table 3.4 indicate that SMLE has improved efficiency of IQMLE by about 50%, and there is no significant difference between IQMLE and QMLE.

In fact, simulation results in Table 3.4 is as expected. In this specific setting, the copula scores are no longer a linear combination of the marginal scores. IQMLE should be less efficient.

3.3.4 Simulations for Robust SMLE

One interesting question that we often encounter in practice is that there could be a bias in marginal parameter estimates if we assume a wrong copula. Compared to an incorrectly specified copula-based likelihood model (Pseudo-MLE), which

estimator is more robust, SMLE or IQMLE ?

Assume the DGP has two exponential marginal distributions with $\mu_1 = \mu_2 = 0.5$, and the dependence is modelled by Plackett, Frank, and Gaussian copula. The parameters in the true copula are set such that there is strong negative dependence between cross-sections. We estimate the marginal parameter by Full-MLE, QMLE, IQMLE, SMLE, and Pseudo-MLE which assumes different copulas other than the true copula.

The simulation results are summarized in Table 3.5, 3.6 and 3.7. There are several interesting observations from this simulation. First, SMLE is relatively more efficient than IQMLE or QMLE. This can be explained by the nonlinear relationship between copula scores and marginal scores. Second, SMLE is relatively more efficient than the group of Pseudo-MLEs. For example, Table 3.5 shows that SMLE is efficient if the assumed copula is Gaussian, Gumbel, Joe, Clayton, FGM, or AMH copulas. We can observe similar patterns in Table 3.6 and 3.7. This simulation results suggest that SMLE is more robust not only than IQMLE, but also than some Pseudo-MLE if the assumed parametric copula is very different from the true one.

3.4 Concluding Remarks

We compare SMLE and IQMLE in a panel setting with information only on the marginal probability distribution.

In practice, SMLE should be more advantageous. Because we usually do not impose the same marginal parameter restrictions, IQMLE is the same as QMLE. In addition, SMLE could not be worse than QMLE since SMLE becomes QMLE if $J_n = 1$. Secondly, as long as there is a nonlinear relationship between copula scores and marginal scores, which is more plausible in practice, SMLE should bring some efficiency gains over IQMLE.

However, the efficiency gains of SMLE over IQMLE come with some computational costs. Usually, the sieve parameter matrix in SMLE is a large sparse matrix.

How efficiently to estimate a large sparse matrix is a key to improve the algorithm of SMLE. This issue has been addressed in Chapter 2.

Most of the theoretical results in this paper are about SMLE when the number of sieve parameters is fixed. In future's research, we should try to compare IQMLE and SMLE directly, ie, when the number of sieve goes to infinity and the sample is very large.

3.5 Appendix: Proofs

Proof of Theorem 3.1. Without loss of generality, assume that we have a panel data with time periods $T = 3$. We also assume (y_{i1}, y_{i2}, y_{i3}) is i.i.d over i, but we allow dependence across time t. Assume the marginal pdf $f_1(y_1; \beta_1), f_2(y_2, ; \beta_2), f_3(y_3; \beta_3)$ are different with distinct parameters. $\beta = (\beta_1', \beta_2', \beta_3')'$

The QMLE is the value of β that maximizes the quasi-likelihood

$$lnL^Q = \sum_i \sum_t lnf_t(y_{it}, \beta)$$

Then the QMLE $\hat{\beta}$ solves the first-order condition

$$\begin{aligned} \sum_i s_i(\hat{\beta}) &= 0 \\ \text{where} \quad s_i(\beta) &= \sum_t s_{it}(\beta) = \sum_t \nabla_\beta lnf_t(y_{it}, \beta) \end{aligned}$$

As a result, QMLE could be regarded as a GMM estimator based on the moment condition

$$\begin{aligned} & Es_i(\beta_0) = 0 \implies \\ & E(\nabla_\beta lnf_1(y_{it}, \beta_{10}) + \nabla_\beta lnf_2(y_{it}, \beta_{20}) + \nabla_\beta lnf_3(y_{it}, \beta_{30})) = 0 \end{aligned} \tag{25}$$

By contrast, IQMLE is a GMM estimator based on the moment

$$E \begin{bmatrix} \nabla_\beta lnf_1(y_1, \beta_1) \\ \nabla_\beta lnf_2(y_2, \beta_2) \\ \nabla_\beta lnf_2(y_3, \beta_3) \end{bmatrix} = 0 \tag{26}$$

If $\beta_1 \neq \beta_2 \neq \beta_3$, the moment condition (25) for QMLE is

$$\begin{aligned} ES_i(\beta) &= E \begin{bmatrix} \frac{\partial f_1(y_{i1}, \beta_1)}{\partial \beta_1} \frac{1}{f_1(y_{i1}, \beta_1)} \\ 0 \\ 0 \end{bmatrix} + E \begin{bmatrix} 0 \\ \frac{\partial f_2(y_{i2}, \beta_2)}{\partial \beta_2} \frac{1}{f_2(y_{i2}, \beta_2)} \\ 0 \end{bmatrix} + E \begin{bmatrix} 0 \\ 0 \\ \frac{\partial f_3(y_{i3}, \beta_3)}{\partial \beta_3} \frac{1}{f_3(y_{i3}, \beta_3)} \end{bmatrix} \\ &= E \begin{bmatrix} \frac{\partial f_1(y_{i1}, \beta_1)}{\partial \beta_1} \frac{1}{f_1(y_{i1}, \beta_1)} \\ \frac{\partial f_2(y_{i2}, \beta_2)}{\partial \beta_2} \frac{1}{f_2(y_{i2}, \beta_2)} \\ \frac{\partial f_3(y_{i3}, \beta_3)}{\partial \beta_3} \frac{1}{f_3(y_{i3}, \beta_3)} \end{bmatrix} = 0 \end{aligned}$$

The moment condition (26) for IQMLE is

$$E\begin{bmatrix}\nabla_\beta lnf_1(y_1,\beta_1)\\ \nabla_\beta lnf_2(y_2,\beta_2)\\ \nabla_\beta lnf_2(y_3,\beta_3)\end{bmatrix} = E\begin{bmatrix}\frac{\partial f_1(y_{i1},\beta_1)}{\partial\beta_1}\frac{1}{f_1(y_{i1},\beta_1)}\\ 0\\ 0\\ 0\\ \frac{\partial f_2(y_{i2},\beta_2)}{\partial\beta_2}\frac{1}{f_2(y_{i2},\beta_2)}\\ 0\\ 0\\ 0\\ \frac{\partial f_3(y_{i3},\beta_3)}{\partial\beta_3}\frac{1}{f_3(y_{i3},\beta_3)}\end{bmatrix} = E\begin{bmatrix}\frac{\partial f_1(y_{i1},\beta_1)}{\partial\beta_1}\frac{1}{f_1(y_{i1},\beta_1)}\\ \frac{\partial f_2(y_{i2},\beta_2)}{\partial\beta_2}\frac{1}{f_2(y_{i2},\beta_2)}\\ \frac{\partial f_3(y_{i3},\beta_3)}{\partial\beta_3}\frac{1}{f_3(y_{i3},\beta_3)}\end{bmatrix} = 0$$

Please note that we disregard 0 because they are redundant moment condition.

We see clearly that both of QMLE and IQMLE is a GMM estimator based on the same moment conditions, so IQMLE is the same as QMLE and they have the same efficiency. Although the proof is based on three periods, it it can be extended easily to more time periods. □

Proof of Theorem 3.2 . See in Prokhorov and Schmidt (2009b) theorem 1. □

Definition 3.2. *A,B,G,K,L N,P,Q ,Z,W,J,E,F are defined as*

$$\begin{aligned}
E(\nabla_\beta lnf_1 \times \nabla'_\beta lnf_1) &= A, \quad & E(\nabla_\beta lnf_1 \times \nabla'_\beta lnf_2) &= G,\\
E(\nabla_\beta lnf_1 \times \nabla'_\beta lnc_s) &= -K, \quad & E(\nabla_\beta lnf_1 \times \nabla'_\gamma lnc_s) &= -P,\\
E(\nabla_\beta lnf_2 \times \nabla'_\beta lnf_2) &= B, \quad & E(\nabla_\beta lnf_2 \times \nabla'_\beta lnc_s) &= -L'\\
E(\nabla_\beta lnf_2 \times \nabla'_\gamma lnc_s) &= -Q', \quad & E(\nabla_\beta lnc_s \times \nabla'_\beta lnc_s) &= N,\\
E(\nabla_\beta lnc_s \times \nabla'_\gamma lnc_s) &= Z, \quad & E(\nabla_\gamma lnc_s \times \nabla'_\gamma lnc_s) &= W,\\
E(\nabla_\beta lnc \times \nabla'_\beta lnc) &= J, \quad & E(\nabla_\beta lnc \times \nabla'_\gamma lnc) &= E,\\
E(\nabla_\gamma lnc \times \nabla'_\gamma lnc) &= F
\end{aligned}$$

The special form of C_{true} has been proved in Prokhorov and Schmidt (2009b) Lemma 1.

Proof of Lemma 3.1. In order to get the asymptotic variance for SMLE(fixed k), we have to decompose $E(\nabla_\beta l_s \nabla'_\beta l_s)$,$E(\nabla_\beta l_s \nabla'_\gamma l_s)$,and $E(\nabla_\theta l_s \nabla'_\gamma l_s)$ Here $l_s = lnf_1 + lnf_2 + lnc_s$, and c_s is the sieve copula.

Given the Definition 3.1 and 3.2, we have

$$\begin{aligned}
E(\nabla_\beta l_s \times \nabla'_\beta l_s) &= E((\nabla_\beta f_1 + \nabla_\beta f_2 + \nabla_\beta lnc_s) \times (\nabla'_\beta f_1 + \nabla'_\beta f_2 + \nabla'_\beta lnc_s)) \\
&= A + G + G' + B - K - L' - K' - L + N \\
E(\nabla_\beta l_s \times \nabla'_\gamma l_s) &= E((\nabla_\beta f_1 + \nabla_\beta f_2 + \nabla_\beta lnc_s) \times \nabla'_\gamma lnc_s) \\
&= -P - Q' + Z \\
E(\nabla_\gamma l_s \times \nabla'_\beta l_s) &= -P' - Q + Z' \\
E(\nabla_\gamma l_s \times \nabla'_\gamma l_s) &= E(\nabla_\gamma lnc_s \times \nabla'_\gamma lnc_s) = W
\end{aligned}$$

We now could write the information matrix of SMLE(fixed k) in an alternative form.

$$\begin{aligned}
I &= \begin{bmatrix} E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\beta}\frac{dl_s(z,\beta_*,\gamma_*)}{d\beta'}] & E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\beta}\frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma'}] \\ E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma}\frac{dl_s(z,\beta_*,\gamma_*)}{d\beta'}] & E[\frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma}\frac{dl_s(z,\beta_*,\gamma_*)}{d\gamma'}] \end{bmatrix} \\
&= \begin{bmatrix} A + G + G' + B - K - L' - K' - L + N & -P - Q' + Z \\ -P' - Q + Z' & W \end{bmatrix}
\end{aligned}$$

□

Proof of Lemma 3.2 . Given Lemma 3.1, the asymptotic variance $\mathbb{V}_{\text{SMLE-K}}$ for SMLE fixed k is the upper left block of the inverse of information matrix I. By partitioned inverse formula, Σ could be expressed as

$$\begin{aligned}
\mathbb{V}_{\text{SMLE-K}} &= [A + G + G' + B - K - L' - K' - L + N \\
&\quad -(-P - Q' + Z)W^{-1}(-P' - Q + Z')]^{-1}
\end{aligned}$$

□

Proof of Lemma 3.3. see Prokhorov and Schmidt (2009b) theorem 4. □

Proof of Theorem 3.3. Given Lemma 3.2 and 3.3, we have the expressions for $\mathbb{V}_{\text{IQMLE}}$ and $\mathbb{V}_{\text{SMLE-K}}$. The IQMLE is as efficient as SMLE (fixed k) if and only if $\mathbb{V}_{\text{SMLE-K}} = \mathbb{V}_{\text{IQMLE}}$, which implies

$$\begin{aligned}
A + G + G' + B - K - L' - K' - L + N - (-P - Q' + Z)W^{-1}(-P' - Q + Z') &= \\
A + B - G' - G + \begin{bmatrix} -G' & -G \end{bmatrix} \begin{bmatrix} A & G \\ G' & B \end{bmatrix}^{-1} \begin{bmatrix} -G \\ -G' \end{bmatrix}
\end{aligned}$$

□

Proof of Lemma 3.4. Given Lemma 3.3

$$\begin{aligned}
\mathbb{V}_{\text{IQMLE}} &= \{A+B-G'-G+\begin{bmatrix}-G' & -G\end{bmatrix}\begin{bmatrix}A & G\\ G' & B\end{bmatrix}^{-1}\begin{bmatrix}-G\\ -G'\end{bmatrix}\}^{-1}\\
&= \{A+G-G+G'+B-G'-G'-G+J-J\\
&\quad +\begin{bmatrix}-G' & -G\end{bmatrix}\begin{bmatrix}A & G\\ G' & B\end{bmatrix}^{-1}\begin{bmatrix}-G\\ -G'\end{bmatrix}\}^{-1}\\
&= \{E(\nabla_\beta l\nabla'_\beta l)-E(\nabla_\beta lnc\nabla'_\beta lnc)+E(\nabla_\beta lncP_{\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}})\nabla'_\beta lnc\}^{-1}\\
&= \{E(\nabla_\beta l\nabla'_\beta l)-E(\nabla_\beta lnc(I-P_{\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}})\nabla'_\beta lnc\}^{-1}\\
&= \{E(\nabla_\beta l\nabla'_\beta l)-E(\nabla_\beta lnc\nabla'_\beta lnc)+E(\widetilde{\nabla_\beta lnc}\widetilde{\nabla'_\beta lnc})\}^{-1}\\
&= \{E(\nabla_\beta l\nabla'_\beta l)-E(ee')\}^{-1}
\end{aligned}$$

where $l = lnf_1 + lnf_2 + lnc$, so l is the true log-likelihood. P_X is the projection matrix of X, e is the regression error of $\nabla_\beta lnc$ on $\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}$, $\widetilde{\nabla_\beta lnc}$ is the fitted value of regression of $\nabla_\beta lnc$ on $\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}$

□

Proof of Lemma 3.5. Given Lemma 3.2, we have

$$\begin{aligned}
\mathbb{V}_{\text{SMLE-K}} &= \{A+G+G'+B-K-L'-K'-L+N\\
&\quad -(-P-Q'+Z)W^{-1}(-P'-Q+Z')\}^{-1}\\
&= \{E(\nabla_\beta l_s\nabla'_\beta l_s)-E(\nabla_\beta l_s\nabla'_{\gamma_N} l_s)E(\nabla_{\gamma_N} l_s\nabla'_{\gamma_N} l_s)^{-1}E(\nabla_{\gamma_N} l_s\nabla'_\beta l_s)\}^{-1}\\
&= \{E(\nabla_\beta l_s\nabla'_\beta l_s)-E(\widehat{\nabla_\beta l_s}\widehat{\nabla'_\beta l_s})\}^{-1}
\end{aligned}$$

where $l_s = lnf_1 + lnf_2 + lnc_s$ and $\widehat{\nabla_\beta l_s}$ is the fitted value of regression of $\nabla_\beta l_s$ on $\nabla'_{\gamma_N} l_s$ □

Proof of Theorem 3.4. It follows easily by Theorem 3.3, and Lemma 3.4 and 3.5 □

Proof of Theorem 3.5. Assume that the sieve likelihood l_s is the same as the true likelihood l, then sieve MLE with fixed k is the same as Full MLE. By Theorem 3.4, we have IQMLE is as efficient as FMLE if and only if

$$\begin{aligned}
E(\nabla_\beta l\nabla'_\beta l)-E(ee') &= E(\nabla_\beta l\nabla'_\beta l)-E(\widehat{\nabla_\beta l}\widehat{\nabla'_\beta l})\\
\iff E(ee') &= E(\widehat{\nabla_\beta l}\widehat{\nabla'_\beta l})\\
\iff E(\nabla_\beta lnc(I-P_{\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}})\nabla'_\beta lnc) &= E(\nabla_\beta lP_{\nabla'_\gamma lnc}\nabla'_\beta l) \qquad (27)
\end{aligned}$$

Under FMLE, $\nabla_\beta lnf_1$ and $\nabla_\beta lnf_2$ are orthogonal to $\nabla_\gamma lnc$.

So,

$$E(\nabla_\beta l P_{\nabla'_\gamma lnc}\nabla'_\beta l) = E(\nabla_\beta lncP_{\nabla'_\gamma lnc}\nabla'_\beta lnc) \tag{28}$$

combine equation 27 and 28, we have

$$\begin{aligned} E(\nabla_\beta lnc(I - P_{\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2\end{bmatrix}})\nabla'_\beta lnc) &= E(\nabla_\beta lncP_{\nabla'_\gamma ln'c}\nabla'_\beta lnc) \\ \iff E(\nabla_\beta lnc\nabla'_\beta lnc) &= E(\nabla_\beta lncP_{\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2 & \nabla'_\gamma lnc\end{bmatrix}}\nabla'_\beta lnc) \\ \iff E(\nabla_\beta lnc\nabla'_\beta lnc) &= E(\nabla_\beta lnc^*\nabla'_\beta lnc^*) \end{aligned} \tag{29}$$

where P_X is the projection matrix of X, and $\nabla_\beta lnc^*$ is the fitted value of regression of $\nabla_\beta lnc$ on $\begin{bmatrix}\nabla'_\beta lnf_1 & \nabla'_\beta lnf_2 & \nabla\gamma' lnc\end{bmatrix}$.

Equation 29 is satisfied when $\nabla_\beta lnc$ is a linear combination of $\nabla_\beta lnf_1$, $\nabla_\beta lnf_2$ and $\nabla\gamma lnc$.

□

Proof of Theorem 3.6. In Example 3.1, SMLE is the same as FMLE. So SMLE has the same asymptotic variance as FMLE. SMLE(FMLE) From asymptotic theory for MLE, we have

$$(\hat{\mu} - \mu_0, \hat{\gamma} - \gamma_0) \to N(0, I^{-1})$$

$$I = E\left[\begin{bmatrix}\nabla_\mu lnf \\ \nabla_\gamma lnf\end{bmatrix}\begin{bmatrix}\nabla_\mu lnf & \nabla_\gamma lnf\end{bmatrix}\right]$$

$$lnf = 1_{11}ln\gamma + (1_{10} + 1_{01})ln(1 - \mu - \gamma) + 1_{00}ln(2\mu - 1 + \gamma)$$

where I is the fisher information matrix, lnf is individual log-likelihood, 1_{11} is the indicator function for $x = x_1$ and $y = y_1$. The asymptotic variance for $\hat{\mu}$ is the upper left block of I^{-1}

From information matrix equality, we know

$$I(\theta)_{i,j} = -E(\frac{\partial^2}{\partial\theta_i\partial\theta_j}lnf(x;\theta)|\theta)$$

We will use information matrix equality to get I^{-1}.

$$\begin{aligned}
\frac{\partial lnf}{\partial \mu} &= -\frac{1_{01}+1_{01}}{1-\mu-\gamma}+\frac{21_{00}}{2\mu-1+\gamma} \\
\frac{\partial lnf}{\partial \gamma} &= \frac{1_{11}}{\gamma}-\frac{1_{01}+1_{10}}{1-\mu-\gamma}+\frac{1_{00}}{2\mu-1+\gamma} \\
\frac{\partial^2 lnf}{\partial \mu^2} &= -\frac{1_{01}+1_{10}}{(1-\mu-\gamma)^2}-\frac{41_{00}}{(2\mu-1+\gamma)^2} \\
\frac{\partial^2 lnf}{\partial \gamma^2} &= -\frac{1_{11}}{\gamma^2}-\frac{1_{01}+1_{10}}{(1-\mu-\gamma)^2}-\frac{1_{00}}{(2\mu-1+\gamma)^2} \\
\frac{\partial^2 lnf}{\partial \mu \partial \gamma} &= -\frac{1_{01}+1_{10}}{(1-\mu-\gamma)^2}-\frac{21_{00}}{(2\mu-1+\gamma)^2} \\
\frac{\partial^2 lnf}{\partial \gamma \partial \mu} &= \frac{\partial^2 lnf}{\partial \mu \partial \gamma}
\end{aligned}$$

We then can get information matrix by taking expectations.

$$\begin{aligned}
I_{11} &= -E(\frac{\partial^2 lnf}{\partial \mu^2}) = \frac{2(1-\mu-\gamma)}{(1-\mu-\gamma)^2}+\frac{4(2\mu-1+\gamma)}{(2\mu-1+\gamma)^2} \\
&= \frac{2(1+\gamma)}{(1-\mu-\gamma)(2\mu-1+\gamma)} \\
I_{22} &= -E(\frac{\partial^2 lnf}{\partial \gamma^2}) = \frac{1}{\gamma}+\frac{2}{1-\mu-\gamma}+\frac{1}{2\mu-1+\gamma} \\
&= \frac{3\mu-1+\gamma-2\mu^2}{(1-\mu-\gamma)(2\mu-1+\gamma)\gamma} \\
I_{12} &= I_{21} = -E(\frac{\partial^2 lnf}{\partial \mu \partial \gamma}) = \frac{2}{1-\mu-\gamma}+\frac{2}{2\mu-1+\gamma} \\
&= \frac{2\mu}{(1-\mu-\gamma)(2\mu-1+\gamma)}
\end{aligned}$$

By partitioned matrix inverse formula, the upper left block of I^{-1} is $(I_{11}-I_{12}I_{22}^{-1}I_{21})^{-1}$.

Plug I_{11},I_{12},I_{21} and I_{22} into the partitioned matrix inverse formula. we get the asymptotic variance of marginal parameters in SMLE. The asymptotic variance $\mathbb{V}_{\text{SMLE}}$ for $\hat{\mu}$

$$\mathbb{V}_{\text{SMLE}} = \frac{\gamma+3\mu-1-2\mu^2}{2} \tag{30}$$

IQMLE is the optimal GMM based on moment condition 31

$$E\begin{bmatrix}\mu 1_{x_1}-(1-\mu)1_{x_0} \\ \mu 1_{y_1}-(1-\mu)1_{y_0}\end{bmatrix}=0 \tag{31}$$

The asymptotic variance $\mathbb{V}_{\text{IQMLE}}$ of $\tilde{\mu}$ for IQMLE is $(D'C^{-1}D)^{-1}$,where D is the derivative

of moment conditions, and C is the covariance matrix for moments.

$$\begin{aligned} C &= \begin{bmatrix} \mu(1-\mu) & \gamma+(1-\mu)^2 \\ \gamma+(1-\mu)^2 & \mu(1-\mu) \end{bmatrix} \\ D &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

The asymptotic variance $\mathbb{V}_{\text{IQMLE}}$ for $\tilde{\mu}$ is

$$\mathbb{V}_{\text{IQMLE}} = \frac{\gamma+1-\mu}{2}$$

More precisely,

$$\begin{aligned} \mathbb{V}_{\text{IQMLE}}^{-1} &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \mu(1-\mu) & \gamma+(1-\mu)^2 \\ \gamma+(1-\mu)^2 & \mu(1-\mu) \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{det(C)}[2\mu(1-\mu)-2\gamma-2(1-\mu)^2] \end{aligned}$$

So,

$$\begin{aligned} \mathbb{V}_{\text{IQMLE}} &= \frac{det(C)}{2\mu(1-\mu)-2\gamma-2(1-\mu)^2} \\ &= \frac{[\mu(1-\mu)]^2-[\gamma+(1-\mu)^2]^2}{2\mu(1-\mu)-2\gamma-2(1-\mu)^2} \\ &= \frac{\mu(1-\mu)+\gamma+(1-\mu)^2}{2} \\ &= \frac{1+\gamma-\mu}{2} \end{aligned}$$

Given the asymptotic variance of SMLE and IQMLE, we could compare them directly.

$$\begin{aligned} \mathbb{V}_{\text{IQMLE}} - \mathbb{V}_{\text{SMLE}} &= \frac{1+\gamma-\mu-(\gamma+3\mu-1-2\mu^2)}{2} \\ &= \frac{2(\mu-1)^2}{2} \geq 0 \end{aligned}$$

So , SMLE is efficient relative to IQMLE in bivariate bernoulli example. □

3.6 Appendix: Tables

Table 3.2: Same Marginal, Distinct Parameters

$\mu_1 = 0.5$	QMLE	IQMLE	SMLE	FMLE
mean	0.499	0.499	0.493	0.500
var(10^{-4})	2.398	2.397	0.712	0.340
bias(10^{-5})	0.002	0.003	4.721	0.001
MSE(10^{-4})	2.395	2.395	1.183	0.340
RE	-	0.999	0.291	0.149
RMSE	-	0.999	0.494	0.149
$\mu_2 = 0.5$	QMLE	IQMLE	SMLE	FMLE
mean	0.500	0.500	0.493	0.499
var(10^{-4})	2.400	2.390	0.729	0.345
bias(10^{-5})	0.001	0.001	4.701	0.002
MSE(10^{-4})	2.401	2.391	1.199	0.345
RE	-	0.999	0.303	0.143
RMSE	-	0.999	0.499	0.143

Table 3.3: Normal Marginals, Gaussian Copula

$J_n = 1$	QMLE	IQMLE	SMLE	FMLE
mean	5	5	5	5
var(10^{-5})	5.023	5.016	5.023	5.025
bias(10^{-10})	3.642	4.803	3.637	3.584
MSE(10^{-5})	5.018	5.011	5.018	5.019
RE	-	0.999	1	1
RMSE	-	0.998	1	1
$J_n = 2$	QMLE	IQMLE	SMLE	FMLE
mean	5	5	5	5
var(10^{-5})	4.805	4.801	4.856	4.803
bias(10^{-11})	9.273	3.051	3.764	9.435
MSE(10^{-5})	4.805	4.801	4.856	4.803
RE	-	0.999	1	0.999
RMSE	-	0.999	1	0.999

Table 3.4: Normal Marginals, Frank Copula

$Jn = 5$	QMLE	IQMLE	SMLE	FMLE
mean	4.9999	4.9999	4.9999	5
var(10^{-5})	1.543	1.544	0.811	0.519
bias(10^{-8})	1.581	1.495	0.851	0.001
MSE(10^{-5})	1.543	1.544	0.811	0.519
RE	-	1	0.525	0.336
RMSE	-	1	0.525	0.336

Table 3.5: Exponential Marginals, Plackett Copula

	MEAN	VAR(10^{-5})	MSE(10^{-5})	RE	RMSE
t	0.500	0.145	0.145	0.032	0.032
frank	0.500	0.155	0.158	0.035	0.035
gaussian	0.499	0.409	0.410	0.092	0.092
gumbel	0.5	4.443	4.438	0.999	0.999
joe	0.5	4.447	4.443	1.000	1.000
clayton	0.500	4.446	4.441	1.000	1.000
fgm	0.486	2.387	20.964	0.537	4.722
amh	0.465	2.237	12.076	0.503	27.202
QMLE	0.5	4.443	4.439	-	-
IQMLE	0.499	4.460	4.467	1.003	1.006
SMLE	0.494	0.313	3.455	0.070	0.778
FMLE	0.499	0.118	0.118	0.026	0.026

Table 3.6: Exponential Marginals, Frank Copula

	MEAN	VAR(10^{-4})	MSE(10^{-4})	RE	RMSE
plackett	0.500	0.899	0.901	0.180	0.180
t	0.501	1.417	1.426	0.284	0.286
gaussian	0.501	2.298	2.301	0.461	0.462
amh	0.465	3.771	15.913	0.756	3.193
fgm	0.486	4.196	6.094	0.841	1.223
clayton	0.500	4.988	4.984	0.999	0.999
gumbel	0.500	4.987	4.983	0.999	0.999
joe	0.500	4.986	4.982	0.999	0.999
QMLE	0.500	4.989	4.984	-	-
IQMLE	0.500	4.987	4.982	0.999	0.999
SMLE	0.492	1.749	2.274	0.351	0.456
FMLE	0.500	0.749	0.749	0.150	0.150

Table 3.7: Exponential Marginals, Gaussian Copula

	MEAN	VAR(10^{-4})	MSE(10^{-4})	RE	RMSE
t	0.499	1.319	1.318	0.268	0.268
plackett	0.500	1.674	1.673	0.340	0.340
frank	0.501	2.053	2.061	0.417	0.419
amh	0.466	3.723	14.659	0.756	2.980
fgm	0.486	4.107	5.873	0.834	1.194
clayton	0.499	4.920	4.917	0.999	0.999
gumbel	0.499	4.919	4.917	0.999	0.999
joe	0.499	4.919	4.916	0.999	0.999
QMLE	0.499	4.921	4.918	-	-
IQMLE	0.499	4.919	4.917	0.999	0.999
SMLE	0.494	2.503	2.794	0.508	0.568
FMLE	0.499	1.311	1.310	0.266	0.266

Bibliography

Ackerberg, D., X. Chen, and J. Hahn (2009): "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," *UCLA Working Paper*.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Amsler, C., A. Prokhorov, and P. Schmidt (2014): "Using copulas to model time dependence in stochastic frontier models," *Econometric Reviews*, 33(5-6), 497–522.

Angrist, J. D., and A. B. Krueger (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87(418), 328–336.

Arellano, M., and C. Meghir (1992): "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets," *The Review of Economic Studies*, 59(3), 537–559.

Aydemir, A., W. Chen, and M. Corak (2009): "Intergenerational Earnings Mobility Among the Children of Canadian Immigrants," *Review of Economics and Statistics*, 91, 377–397.

Bhattacharya, D., and B. Mazumder (2011): "A Nonparametric Analysis of Black-White Differences in Intergenerational Income Mobility in the United States," *Quantitative Economics*, 2, 335–379.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 37(4), 1705–1732.

Birgé, L., and P. Massart (1997): *From model selection to adaptive estimation*. Springer.

Björklund, A., and M. Jäntti (1997): "Intergenerational Income Mobility in Sweden Compared to the United States," *The American Economic Review*, 87(5), 1009–1018.

Black, S. E., and P. J. Devereux (2011): "Recent Developments in Intergenerational Mobility," in *Handbook of Labor Economics*, vol. 4B, pp. 1487–1541. Elsevier.

Bouezmarni, T., and J. V. K. Rombouts (2010): "Nonparametric density estimation for multivariate bounded data," *Journal of Statistical Planning and Inference*, 140(1), 139–152.

Bratsberg, B., K. Røed, O. Raaum, R. Naylor, M. Jäntti, T. Eriksson, and E. Österbacka (2007): "Nonlinearities in Intergenerational Earnings Mobility: Consequences for Cross-Country Comparisons," *The Economic Journal*, 117(519), C72–C92.

Bunea, F., A. B. Tsybakov, M. H. Wegkamp, and A. Barbu (2010): "Spades and mixture models," *The Annals of Statistics*, 38(4), 2525–2558.

Burda, M., and A. Prokhorov (2013): "Copula-Based Factorization for Bayesian Infinite Mixture Models," *Concordia University Working paper*.

Cai, Z., M. Das, H. Xiong, and X. Wu (2006): "Functional Coefficient Instrumental Variables Models," *Journal of Econometrics*, 133(1), 207–241.

Cai, Z., and Q. Li (2008): "Nonparametric Estimation of Varying Coefficient Dynamic Panel Data Models," *Econometric Theory*, 24(05), 1321–1342.

Candes, E., and T. Tao (2007): "The Dantzig Selector: Statistical Estimation When p Is Much Larger than n," *The Annals of Statistics*, 35(6), pp. 2313–2351.

Candes, E. J. (2006): "Modern statistical estimation via oracle inequalities," *Acta Numerica*, 15, 257–325.

Checchi, D., A. Ichino, and A. Rustichini (1999): "More Equal but Less Mobile? Education Financing and Intergenerational Mobility in Italy and in the US," *Journal of Public Economics*, 74, 351–393.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, X., Y. Fan, and V. Tsyrennikov (2006): "Efficient estimation of semiparametric multivariate copula models," *Journal of the American Statistical Association*, 101(475), 1228–1240.

Chen, X., and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152(1), 46–60.

Corak, M. (2013): "Income Inequality, Equality of Opportunity, and Intergenerational Mobility," *Journal of Economic Perspectives*, 27(3), 79–102.

Corak, M., and A. Heisz (1999): "The Intergenerational Earnings and Income Mobility of Canadian Men: Evidence from Longitudinal Income Tax Data," *The Journal of Human Resources*, 34(3), 504–533.

Corcoran, M., R. Gordon, D. Laren, and G. Solon (1992): "The Association between Men's Economic Status and Their Family and Community Origins," *Journal of Human Resources*, 27, 575–601.

Datcher, L. P. (1982): "Effects of Community and Family Background on Achievement," *The Review of Economics and Statistics*, 64, 32–41.

Devroye, L., and G. Lugosi (2000): *Combinatorial Methods in Density Estimation*. Springer.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

Fisher, G., and M.-C. Voia (2002): "Estimating Systems of Stochastic Coefficients Regressions When Some of the Observations Are Missing," *Handbook of applied econometrics and statistical inference*.

Frees, E. W., and P. Wang (2005): "Credibility using copulas," *North American Actuarial Journal*, 9(2), 31–48.

Ghosal, S. (2001): "Convergence rates for density estimation with Bernstein polynomials," *Annals of Statistics*, 29(5), 1264–1280.

Gottschalk, P., and T. M. Smeeding (1997): "Cross-National Comparisons of Earnings and Income Inequality," *Journal of Economic Literature*, 32, 633–686.

Gustafsson, B. (1994): "The Degree and Pattern of Income Immobility in Sweden," *Review of Income and Wealth*, 40, 67–86.

Hall, P. G., and J. S. Racine (2013): "Infinite Order Cross-Validated Local Polynomial Regression," Discussion paper.

Heyde, C. C. (1997): *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. Springer.

Hill, C., and F. Stafford (1974): "Allocation of Time to Preschool Children and Education Opportunity," *Journal of Human Resources*, 9, 323–341.

Hong, Y., and T.-H. Lee (2003): "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models," *Review of Economics and Statistics*, 85(4), 1048–1062.

Inoue, A., and G. Solon (2010): "Two-Sample Instrumental Variables Estimators," *Review of Economics and Statistics*, 92(3), 557–561.

James, G. M., and P. Radchenko (2009): "A generalized Dantzig selector with shrinkage tuning," *Biometrika*, 96(2), 323–337.

Jantti, M., B. Bratsberg, K. Roed, O. Raaum, R. Naylor, E. Osterbacka, A. Bjorklund, and T. Eriksson (2006): "American Exceptionalism in a New Light: A Comparison of Intergenerational Earnings Mobility in the Nordic Countries, the United Kingdom and the United States," *IZA Discussion Papers 1938*.

Joe, H. (2005): "Asymptotic efficiency of the two-stage estimation method for copula-based models," *Journal of Multivariate Analysis*, 94(2), 401–419.

Juhl, T. (2005): "Functional-coefficient Models under Unit Root Behaviour," *The Econometrics Journal*, 8(2), 197–213.

Kiker, B., and C. M. Condon (1981): "The influence of socioeconomic background on the earnings of young men," *The Journal of Human Resources*, 16(1), 94–105.

Koltchinskii, V. (2009): "The Dantzig selector and sparsity oracle inequalities," *Bernoulli*, 15(3), 799–828.

Lefranc, A., and A. Trannoy (2005): "Intergenerational Earnings Mobility in France: Is France more Mobile than the US?," *Annales d'Economie et de Statistiques*, 78, 57–75.

Levene, H. (1960): "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. by I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann, pp. 278–292. Stanford University Press, Menlo Park, CA.

Lorentz, G. (1986): *Bernstein Polynomials*. University of Toronto Press.

Murtazashvili, I. (2012): "An Alternative Measure of Intergenerational Income Mobility Based on a Random Coefficient Model," *Journal of Applied Econometrics*, 27, 1161–1173.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Österberg, T. (2000): "Intergenerational Income Mobility in Sweden: What Do Tax-data Show?," *Review of Income and Wealth*, 46(4), 421–436.

Panchenko, V., and A. Prokhorov (2013): "Efficient Estimation of Parameters in Marginals," *Concordia University Working Paper*.

Peters, H. (1992): "Patterns of intergenerational mobility in income and earnings," *Review of Economics and Statistics*, 74, 456–466.

Pfeffermann, D. (1984): "On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients," *Journal of the Royal Statistical Society. Series B. Methodological*, 46(1), 139–148.

Pfeifer, D., D. Strassburger, and J. Philipps (2009): "Modelling and simulation of dependence structures in nonlife insurance with Bernstein copulas," in *39th International ASTIN Colloquium, Helsinki*.

Prokhorov, A., and P. Schmidt (2009a): "GMM redundancy results for general missing data problems," *Journal of Econometrics*, 151(1), 47–55.

——— (2009b): "Likelihood-based estimation in a panel setting: robustness, redundancy and validity of copulas," *Journal of Econometrics*, 153(1), 93–104.

Sancetta, A. (2007): "Nonparametric estimation of distributions with given marginals via Bernstein-Kantorovich polynomials: L1 and pointwise convergence theory," *Journal of Multivariate Analysis*, 98(7), 1376–1390.

Sancetta, A., and S. Satchell (2004): "The Bernstein Copula And Its Applications To Modeling And Approximations Of Multivariate Distributions," *Econometric Theory*, 20(03), 535–562.

Sewell, W. H., and R. M. Hauser (1975): *Education, Occupation, and Earnings: Achievement in the Early Career*. Academic Press, New York.

Shen, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

Shen, X., and W. H. Wong (1994): "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22(2), 580–615.

Sklar, A. (1959): "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Solon, G. (1992): "Intergenerational Income Mobility in the United States," *The American Economic Review*, 82(3), 393–408.

——— (1999): "Intergenerational Mobility in the Labor Market," in *Handbook of Labor Economics*, vol. Volume 3, Part A, pp. 1761–1800. North Holland.

——— (2002): "Cross-Country Differences in Intergenerational Earnings Mobility," *The Journal of Economic Perspectives*, 16, 59–66.

Solon, G. (2004): "A Model of Intergenerational Mobility Variation over Time and Place," in *Generational Income Mobility in North America and Europe*, ed. by M. Corak, pp. 38–47. Cambridge University Press, Cambridge.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 187–195.

Su, L., I. Murtazashvili, and A. Ullah (2013): "Local Linear GMM Estimation of Functional Coefficient IV Models with an Application to Estimating the Rate of Return to Schooling," *Journal of Business and Economic Statistics*, forthcoming.

Tenbusch, A. (1994): "Two-dimensional Bernstein polynomial density estimators," *Metrika*, 41(1), 233–253, Metrika.

Vitale, R. (1975): "A Bernstein polynomial approach to density function estimation," in *Statistical inference and related topics*, ed. by M. Puri. New York: Academic Press.

Zhang, W., S.-Y. Lee, and X. Song (2002): "Local Polynomial Fitting in Semivarying Coefficient Model," *Journal of Multivariate Analysis*, 82(1), 166–188.

Zheng, Y. (2011): "Shape restriction of the multi-dimensional Bernstein prior for density functions," *Statistics and Probability Letters*, 81(6), 647–651.

Zimmerman, D. (1992): "Regression toward Mediocrity in Economic Stature," *American Economic Review*, 82, 409–429.

Appendix A

CopulasToolbox: A Matlab Package For Copulas Modeling

The idea of creating *CopulasToolbox* package in Matlab comes from my doctoral thesis. In a panel data setting, given that the true marginal probability distribution is known, we want to estimate the marginal parameters when using a copula to capture the dependence between cross-sections. However, we usually do not know the true copula that underlies the data generating process, so specifying a wrong copula may result in a biased estimation.

I wish to conduct a simulation-based study of the bias caused by an incorrectly specified copula. However, there are two main difficulties with such simulations. First, since there are many different univariate marginal probability distributions and many copula functions, I will have a large number of combinations of marginals and copulas. It is tedious to modify the code each time I have a different data generating process (DGP). Second, even if I have a flexible platform for multivariate analysis, running simulations are very time consuming. Normally, in a dual-core computer, it can take several hours to run a single simulation. As I have to consider many different cases, I need to run different simulations many times.

Facing the obstacles above, I wish to create an efficient toolbox general enough to handle different marginals and copulas. The first difficulty can be solved by objective-oriented programming (OOP). I just need to create three classes, one for marginal distributions, one for copula distributions and one for multivariate probability distributions. One combination of marginals and copulas is just an entity in the class of multivariate probability distribution. OOP is supported in many languages such as C++, Java, Matlab, R, and Python. The second difficulty can be solved by parallel computing on a cluster of computers. Fortunately, Concordia University

have a cluster called Cirrus[1], on which Matlab can be run in parallel.

Although *CopulasToolbox* is usually used on a cluster, it can also be used in a normal personal computer. Instead of using simulated data, we can also use empirical data.

This manual is organized as follows. Section A.1 presents three main classes, with their class properties and methods. Section A.2 will give some simulation examples in multivariate probability modelling. Section A.3 discusses possible further developments.

A.1 Three Classes

By Sklar's (1959) theorem, a copula is a multivariate distribution function that connects two or more marginal distributions to form a joint multivariate distribution. More precisely,

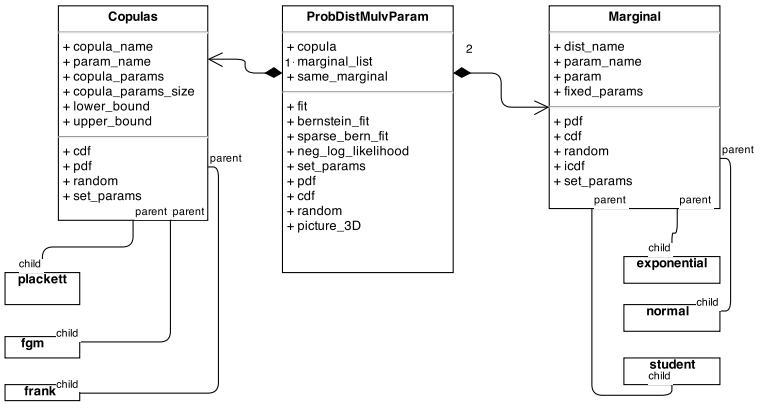
Theorem A.1. *(Sklar, 1959, p229-230). Let H be a T-dimensional distribution function with marginals $F_1, \ldots, F_T$. Then there exists a T-dimensional copula C such that for all $(y_1, \ldots, y_T)$*

$$H(y_1, \ldots, y_T) = C(F_1(y_1), \ldots, F_T(y_T)). \tag{1}$$

If $F_1, \ldots, F_T$ are continuous, then C is unique. Conversely, if C is a T-dimensional copula and $F_1, \ldots, F_T$ are distribution functions, then the function H in (1) is a T-dimensional distribution function with marginals $F_1, \ldots, F_T$

By Theorem A.1, it appears natural that a multivariate distribution class can be constructed based on a copula distribution class and a marginal distribution class. Figure A.1 presents the design of *CopulasToolbox*.

Figure A.1: Design of CopulasToolbox



[1]see link: https://www.encs.concordia.ca/aits/public/sag/systems/cirrus/

A.1.1 Copulas Class

Copulas Class is a general description for different copulas. For now, this class includes 11 different copulas. They are Gaussian, t, Gumbel, Clayton, Frank, Plackett, Ali-Mikhail-Haq, Farlie-Gumbel-Morgenstern , Joe, Logistic, and Bernstein copulas. Each copula distribution has its proper probability density function, cumulative density function, and random number generation function.

Methods It has four main methods: pdf, cdf, random, and Copulas

Table A.1: Copulas Class: Methods

Names	description
Copulas	constructor
pdf	copula probability density function
cdf	copula cumulative density function
random	bivariate copula random number generation

It is very easy to construct a copula, you just to need to provide the name and parameter of the copula. For example, the following code generates the Plackett copula.

Example A.1. *Construction of Plackett copula*

```
CopulaName='plackett';
CopulaParams={0.039};
plackett=Copulas(CopulaName,CopulaParams);
```

The Copulas class supports multi-parameters copulas. For example, we want to construct t copula with correlation matrix ρ and degree of freedom ν.

Example A.2. *Construction of t copula with correlation matrix ρ and degree of freedom ν.*

```
CopulaName='t';
rho=[1 0.5 ; 0.5 1];
nu=3;
CopulaParams={rho,nu};
t=Copulas(CopulaName,CopulaParams);
```

This copulas class is easily extensible.

Example A.3. *Add a new copula into existing Copulas class.*

We want to add a new copula called 'ind'. It is the product of two marginal distributions. It takes two steps to add this new copula.

First, define functions pdf, cdf, statistics,and random for ind copula in the folder "./tools/-copulas/ind". Let's say, these three functions are named as ind_cdf.m, ind_pdf.m, ind_rnd.m and ind_stats.m.

Second, in the file '/@Copulas/private/getCopula.m', we add the following lines in the sub-function 'getBuiltinCopula'

```
\%  independent copula
j=1;
s(j).name='independent';
s(j).code='ind';
s(j).paramName = {'theta'};
s(j).paramNum = 1;
s(j).cdffunc = @ind_cdf ;
s(j).pdffunc = @ind_pdf;
s(j).randfunc = @ind_rnd;
s(j).statsfunc=@ind_stats;
s(j).lowerBound={-Inf};
s(j).upperBound={Inf};
```

Now, 'ind' copula can be constructed by the Copulas class.

```
CopulaName='ind';
CopulaParams={0};
ind=Copulas(CopulaName,CopulaParams);
```

A.1.2 Marginal Distribution Class

Marginal distribution class is a subclass of ProbDistUnivParam, which is a Matlab built-in class for parametric univariate probability distributions. So, the class Marginal inherits all the properties and methods defined in ProbDistUnivParam. However, Marginal class also contains some additional properties and methods.

Before listing the properties and methods, let us consider an example of creating a Marginal distribution .

Example A.4. *Creating Exponential marginal distribution.*

```
m1Name='exp';
m1Params=[0.5];
m1=Marginal(m1Name,m1Params);
m1.FixedParams=[0];
```

So the function Marginal() is the constructor of a Marginal distribution. FixedParams , as a property of Marginal class, indicate whether we fix the parameter in maximum likelihood estimation. Let's consider an example.

Example A.5. *Normal marginals with mean 0 and variance 0.5. The variance is fixed at 0.5.*

```
m1Name='norm';
m1Params=[0 , 0.5];
m1=Marginal(m1Name,m1Params);
m1.FixedParams=[0 1];
```

Suppose the variance is known, we want to estimate the mean parameter by MLE.

Table A.2: Marginal Class: Methods

Names	description
Marginal	constructor
cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean
median	Median
paramci	Confidence intervals for the parameters
pdf	Probability density function
random	Random number generation
std	Standard deviation
var	Variance

Table A.3: Marginal Class: Properties

Names	description
DistName	name of the distribution
InputData	structure containing data used to fit the distribution
NLogL	negative log likelihood for fitted data
NumParams	number of parameters
ParamNames	cell array of NumParams parameter names
Params	array of NumParams parameter values
FixedParams	logical vector indicating which parameters are fixed rather than estimated
ParamDescription	cell array of NumParams strings describing the parameters
ParamCov	covariance matrix of parameter values
Support	structure describing the support of the distribution

A.1.3 ProbDistMulvParam Class

ProbDistMulvParam is a class for parametric multivariate probability densities. By Theorem A.1, we see that a given copula and marginal distributions imply a joint probability density. So naturally, ProbDistMulvParam is based on Marginal class and Copulas Class. Let's see this in an example.

Example A.6. *Construction of object of ProbDistMulvParam class, which has Plackett as copula and two exponential marginals.*

```
%% %%%%%%%%%%%%%%%%%%%%%%
%1. Marignal distribution%
%%%%%%%%%%%%%%%%%%%%%%%%
%first marginal
m1Name='exp';
m1Params=[0.5];
m1=Marginal(m1Name,m1Params);
m1.FixedParams=[0];

%second marginal
m2Name='exp';
m2Params=[0.2];
```

```
m2=Marginal(m1Name,m1Params);
m2.FixedParams=[0];

%% %%%%%%%%%%%%%%%%%%%%%
%2.Copula distribution %
%%%%%%%%%%%%%%%%%%%%%%%%
%true copula
trueCopulaName='plackett';
trueCopulaParams={3};
trueCopula=Copulas(trueCopulaName,trueCopulaParams);

%% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  Multivariate Probability Distribution %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
trueProbDist=ProbDistMulvParam(trueCopula,{m1,m2});
```

Table A.4: ProbDistMulvPram Class: Methods

Names	description
ProbDistMulvParam	constructor
fit	Given the data, using MLE to estimate parameters in marginals and copula distribution
bernsteinFit	Given the data, using Sieve MLE to estimate parameters in marginals and berstein copula
NlogLikelihood	negative loglikelihood given data set
random	random number generation
setParams	set new parameters for marginals and copulas
pdf	probability density function
cdf	cumulative probability density function
picture3D	provide data for a 3D surface for the whole panel

Table A.5: ProbDistMulvParam Class: Properties

Names	description
copula	copula probability distribution
marginals	marginal probability distribution
InputData	input data
SameMarginal	logical indicate whether two marginal are considered the same in MLE.

SameMarginal is worth a note. This property is set to be False by default. However, if we have a prior knowledge that both marginals are the same, SameMarginal can be set explicitly as True.

A.2 Examples with Simulations

This section will give examples used in simulations. We will see how to generate data in Example A.7, have it fitted by an assumed multivariate density in Example A.8, and visualize the estimated pdf in a 3-D surface in Example A.9. Example A.10 demonstrates how to fit marginal parameters with a nonparametric copula–Bernstein Copula.

Example A.7. *Bivariate random number generation.*

```
Nobs=1e+3;
trueProbDist=ProbDistMulvParam(trueCopula,{m1,m2});
Y=random(trueProbDist,Nobs);
```

Here, trueCopula , m1 and m2 have been defined in previous example. We can also invoke random function as

```
Y=trueProbDist.random(Nobs);
```

Example A.8. *Fitting copula and marginal with given data.*

```
%% %%%%%%%%%%%%%%%%%%%%%%
%1. Marignal distribution%
%%%%%%%%%%%%%%%%%%%%%%%%%

%first marginal
m1Name='exp';
```

```
m1Params=[0.5];
m1=Marginal(m1Name,m1Params);
m1.FixedParams=[0];

%second marginal
m2Name='exp';
m2Params=[0.5];
m2=Marginal(m2Name,m2Params);
m2.FixedParams=[0];

%% %%%%%%%%%%%%%%%%%%%%%
%2.Copula distribution %
%%%%%%%%%%%%%%%%%%%%%%%%

%true copula
trueCopulaName='plackett';
trueCopulaParams={0.039};
trueCopula=Copulas(trueCopulaName,trueCopulaParams);

%% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  Multivariate Probability Distribution %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

trueProbDist=ProbDistMulvParam(trueCopula,{m1,m2});

%% %%%%%%%%%%%
% simulation %
%%%%%%%%%%%%%%

Nobs=1e+3; % number of observations
x0=[0.4 0.4  0.5]; % starting point for fmincon when using trueProbDist
Y=random(trueProbDist,Nobs); % random number generation
fmle=fit(x0,trueProbDist,Y)

                                          First-order      Norm of
```

```
 Iter F-count            f(x)  Feasibility   optimality         step
    0       4    4.413037e+02    0.000e+00    7.310e+02
    1      13    3.538780e+02    0.000e+00    1.622e+02    3.385e-01
    2      17    2.551253e+02    0.000e+00    4.125e+04    3.163e-01
    3      24    2.034738e+02    0.000e+00    4.182e+04    7.761e-02
    4      29    1.327612e+02    0.000e+00    6.406e+02    1.539e-01
    5      33    1.276413e+02    0.000e+00    6.692e+02    2.547e-02
    6      38    5.872722e+01    0.000e+00    7.850e+02    1.219e-01
    7      42    3.247599e+01    0.000e+00    9.450e+02    2.694e-02
    8      47   -3.219737e+00    0.000e+00    2.283e+03    4.237e-02
    9      52   -3.403438e+01    0.000e+00    3.382e+02    4.428e-02
   10      56   -3.539860e+01    0.000e+00    2.243e+02    2.991e-02
   11      60   -3.584991e+01    0.000e+00    1.795e+01    4.771e-03
   12      64   -3.586809e+01    0.000e+00    2.558e+00    3.360e-03
   13      68   -3.586897e+01    0.000e+00    7.349e-01    6.288e-04
   14      72   -3.586898e+01    0.000e+00    1.263e-01    8.803e-06

Local minimum possible. Constraints satisfied.

fmincon stopped because the size of the current step is less than
the selected value of the step size tolerance and constraints were
satisfied to within the default value of the constraint tolerance.

<stopping criteria details>


fmle=

    0.5075    0.4840    0.0352
```

In this example, we first generate data from a multivariate density with exponential marginals and the dependence is modelled by the Plackett copula with 0.039. Suppose we correctly specify the copula and marginal family, but we do not know the values of the parameters. Then we could fit plackett and marginal to the generated data. The $fmle = [0.5075, 0.4840, 0.0352]$ is the MLE estimates. 0.5075 is estimates for the first exponential marginals, 0.4840 is for the second marginals, and 0.0352 is for the Plackett copulas. We see that MLE estimates are near the true

values $[0.5, 0.5, 0.039]$.

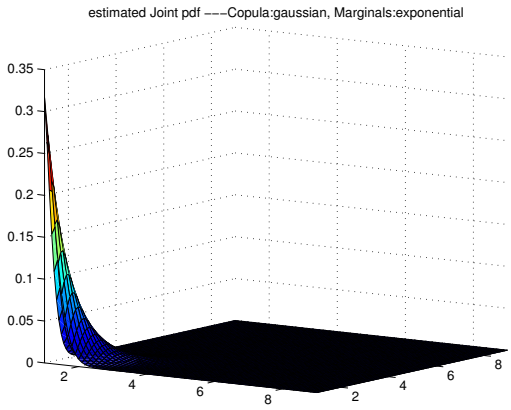
Example A.9. *Visualization of estimated joint pdf and copula pdf by 3-D surface*

```
%% %%%%%%%%%%%%%%%%%%%%%
%1. Marignal distribution%
%%%%%%%%%%%%%%%%%%%%%%%%%

%first marginal
m1Name='exp';
m1Params=[0.5];
m1=Marginal(m1Name,m1Params);
m1.FixedParams=[0];

%second marginal
m2Name='exp';
m2Params=[0.5];
m2=Marginal(m2Name,m2Params);
m2.FixedParams=[0];

%% %%%%%%%%%%%%%%%%%%%%%
%2.Copula distribution %
%%%%%%%%%%%%%%%%%%%%%%%

%true copula
trueCopulaName='gaussian';
trueCopulaParams={0.9};
trueCopula=Copulas(trueCopulaName,trueCopulaParams);

%% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%  Multivariate Probability Distribution %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

trueProbDist=ProbDistMulvParam(trueCopula,{m1,m2});



%% %%%%%%%%%%
```

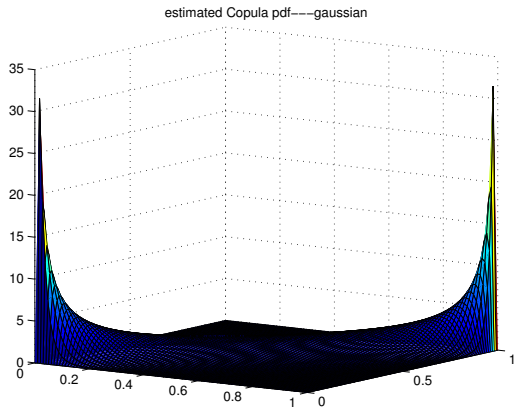
```
% simulation %
%%%%%%%%%%%%%

Nobs=1e+3; % number of observations
x0=[0.4 0.4  0.5]; % starting point for fmincon when using trueProbDist


%random data generation
Y=random(trueProbDist,Nobs);
fmle=fit(x0,trueProbDist,Y);

%% Joint-pdf surface
trueProbDist=setParams(fmle,trueProbDist);
[x1,y1,z1]=picture3D(trueProbDist,1);


figure(1);
surf(x1,y1,z1);
title(sprintf('estimated Joint pdf ---Copula:%s, Marginals:%s', ...
    trueProbDist.Copula.CopulaName,trueProbDist.Marginals{1}.DistName));
xlim([1 9]);
ylim([1 9]);
zlim_j_t=zlim;
view(35,10);

%% Copula pdf surface
[x3,y3,z3]=picture3D(trueProbDist,2);

figure(2);

surfc(x3,y3,z3);
title(sprintf('estimated Copula pdf---%s',trueProbDist.Copula.CopulaName));
zlim_c_t=zlim;
view(35,10);
```

Figure A.2: Visulization: 3D Scatter

(a) Estimated Joint pdf



(b) Estimated Copula pdf



In Example A.9, the data is generated by the Gaussian copula and exponential marginals. Then we estimate the marginal and copula parameters by full maximum likelihood estimation. Finally, the estimated joint probability density function and estimated copula density are visualized in a 3 dimensional surface.

Example A.10. *Fitting marginals to a data with a nonparametric copula–Bernstein Copula.*

```
%% %%%%%%%%%%%%%%%%%%%%%%%%
%1. Marignal distribution%
%%%%%%%%%%%%%%%%%%%%%%%%%%

%first marginal
m1Name='exp';
m1Params=[0.5];
m1=Marginal(m1Name,m1Params);
m1.FixedParams=[0];

%second marginal
m2Name='exp';
m2Params=[0.5];
m2=Marginal(m2Name,m2Params);
m2.FixedParams=[0];
```

```
%% %%%%%%%%%%%%%%%%%%%%%
%2.Copula distribution %
%%%%%%%%%%%%%%%%%%%%%%%

%true copula
trueCopulaName='frank';
trueCopulaParams={39};
trueCopula=Copulas(trueCopulaName,trueCopulaParams);

%bernstein copula
assumedCopulaName='bernstein';
Jn=6;
assumedCopulaParams= {ones(1,Jn^2)+1;Jn};
BernsteinCopula=Copulas(assumedCopulaName,assumedCopulaParams);




%% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                            %
%  Multivariate Probability Distribution %
%                                            %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

trueProbDist=ProbDistMulvParam(trueCopula,{m1,m2});
BernsteinProbDist=ProbDistMulvParam(BernsteinCopula,{m1,m2});

%% %%%%%%%%%%
% simulation %
%%%%%%%%%%%%%

Nobs=1e+3; % number of observations
Nrep=0.01e+3; % number of replications
x0=[0.4 0.4 0.5]; % starting point for fmincon when using trueProbDist
x1=[0.4 0.4];
```

```
%random data generation
Y=random(trueProbDist,Nobs);

%full-MLE
fmle=fit(x0,trueProbDist,Y);
%sieve-MLE
smle=bernsteinFit(x1,Jn,BernsteinProbDist,Y);
```

A.3 Further Development

Many other desirable features of copulas have not been implemented. In addition, some numerical issues have not been elegantly solved yet.

First, *CopulasToolbox* only supports data with 2 dimensions, because the copula random number generation is implemented for bivariate data. It would be good to extend the package to support 3 or more dimensional data. For example, we can consider using compounding construction algorithm.

Second, for some copulas, the random number generation algorithm is not very satisfactory. For example, for the Joe copula, random numbers are generated through solving an equation based on conditional cumulative probability function (ccdf). However, it is possible that for some copulas, the ccdf do not have an explicit form. Then ccdf should be obtained by numerically evaluating the derivative of cdf. But numerical derivatives could be very inaccurate.

Third, MLE and Sieve MLE are implemented by Matlab built-in optimizer: fmincon and fminunc. In simulations, it can be observed that initial starting point for the optimizer can be important. This requires some additional prior knowledge of the data. Furthermore, fmincon and fminunc are slow if the parameter space is large. Especially for sieve MLE, it requires an optimizer to estimate a large sparse matrix. This problem can be solved by using the Dantzig selector to select the non-zeros elements in the Bernstein copula parameter vector, and to fix the zero or near zero elements in the subsequent optimization.

Fourth, automatic differentiation provides an accurate approximation to the derivative of any function written in a computer code, and it is much more stable than traditional numerical differentiation. So we can use this technique to provide the first order derivative, which can greatly improve the computational efficiency.

Finally, goodness-of-fit tests could be included. It is convenient to recommend a best fitting copula to end users.

In the future development, I may consider implementing a similar package in C++ for two reasons. First, C++ is much faster than Matlab, so we may have similar performance on a personal computer in C++ as in Matlab on a cluster. Second, many favourable numerical packages are available. For example, Armadillo is available for linear algebra; CPPAD for automatic differentiation; NLopt for nonlinear optimization; Boost has a general purpose library, which includes a very good package for Probability distributions.