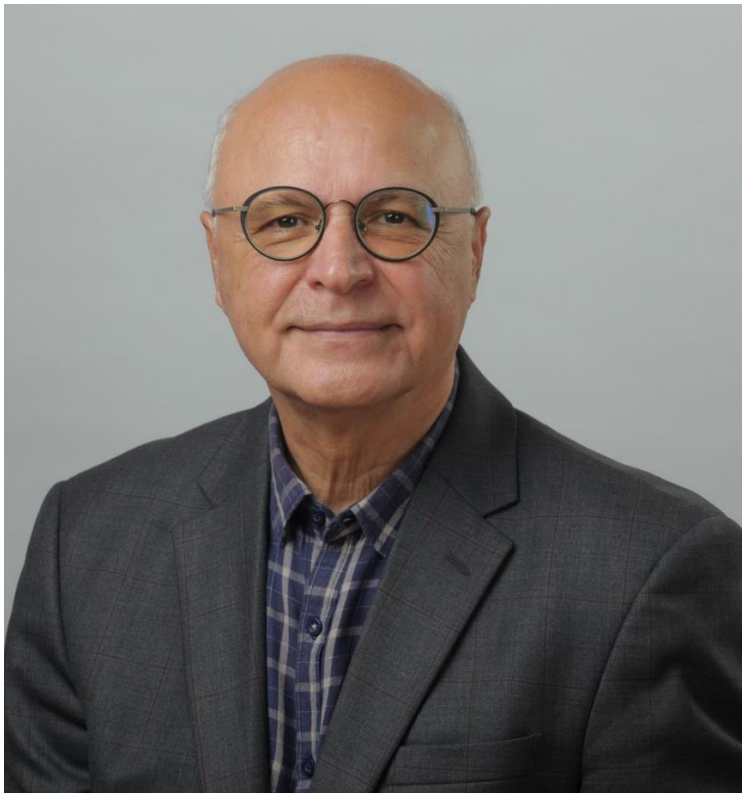
THE ET INTERVIEW:
ESFANDIAR (ESSIE) MAASOUMI

Interviewed by Aman Ullah
University of California, Riverside



Essie Maasoumi—Professor of Economics, Emory University

Essie Maasoumi was born of Iranian parents in Iran, on March 5, 1950. After his early education in Iranian schools he obtained all of his degrees from the London School of Economics (LSE), which include a B.Sc. (1972) in Mathematical Economics & Econometrics, an M.Sc. (1973) in Statistics, and a Ph.D. (1977) in Econometrics under Denis Sargan. During his Ph.D. studies at the LSE he also served as a Lecturer in Economics at the LSE, and then as Lecturer in Econometrics at the University of Birmingham, United Kingdom. Then he moved in 1977 to the University of Southern California and began his long career. After serving at Indiana University and Southern Methodist

University, he joined Emory University, Atlanta, where he has been the Arts and Sciences Distinguished Professor of Economics since 2008.

Essie's research, along with his strong interest in the philosophy and history of science, has focused on many branches of economics and econometrics. These include more than 100 published papers on policy evaluation and treatment effects, financial econometrics and forecasting, multidimensional well-being and poverty, and theoretical econometrics[1]. His earliest work in Econometrica (1978)[2] and Journal of Econometrics (1980)[3] came out from his PhD thesis and focused on uncertain and misspecified models for inference and forecasting. Several strands of thinking and themes originate in that early work and manifest throughout his later work (Maasoumi et al. in Econometric Reviews (2016)[4]): all models are misspecified and therefore require appropriate econometric methodology to analyze them; non-normality and nonlinearity of economic processes should be realized; in empirical work full distribution of outcomes should be considered rather than just the simple functions of these distributions; to incorporate some of these issues special efforts should be placed on test-based shrinkage and combination methods. Essie has utilized and developed frequentist, Bayesian, and Information Theory (IT) methodologies and demonstrated how and why they work or not work.

Essie is a leader in the field of IT, first developed in communication theory. This has allowed him to formulate many popular economics and econometric hypotheses, such as dependence, goodness of fit, predictability, symmetry, and reversibility in their natural forms, and to provide solid tests of them based on entopic functions. In another domain of research on stochastic dominance, his paper with Linton and Whang[5] in The Review of Economic Studies (2005) is an influential and highly cited paper, in which a robust general test of stochastic dominance of arbitrary order is provided for uniform ranking of outcomes. This paper is an outcome of his path breaking work on indices for measuring inequality, multidimensional well-being, poverty, and mobility, subjects in which Essie is one of the leaders. The multidimensional parts (e.g., Massoumi's highly cited Econometrica (1986)[6]) of his work were motivated by the early influence of Nobel Laureate Amartya Sen who was a long-time professor at the LSE. Consistent with his favored theme of the whole distribution of outcomes, Essie has made innovative contributions on program evaluation and

[1] Please see the link http://economics.emory.edu/home/people/faculty/maasoumi-esfandiar.html.

[2] Maasoumi, E. (1978). A modified Stein-like estimator for the reduced form coefficients of simultaneous equations. Econometrica: Journal of the Econometric Society, 695-703.

[3] Maasoumi, E. (1980). A ridge-like method for simultaneous estimation of simultaneous equations. Journal of Econometrics, 12(2), 161-176.

[4] Caner, M., Maasoumi, E., & Riquelme, J. A. (2016). Moment and IV selection approaches: A comparative simulation study. Econometric Reviews, 35(8-10), 1562-1581.

[5] Linton, O., Maasoumi, E., & Whang, Y. J. (2005). Consistent testing for stochastic dominance under general sampling schemes. The Review of Economic Studies, 72(3), 735-765.

[6] Maasoumi, E. (1986). The measurement and decomposition of multi-dimensional inequality. Econometrica: Journal of the Econometric Society, 991-997.

treatment effects with an emphasis on studying heterogeneity of outcomes instead of merely summary measures such as average treatment effect. Using IT based procedures, he and his students, such as Le Wang, have extensively studied the impact of class size, marriage, and union membership on wage outcomes and earnings gaps, which have appeared in leading journals such as the Journal of Political Economy (2019) and Journal of Econometrics (2017) among others. These and all of his research contributions have made a lasting impact on the profession, and this is reflected in the special issue of Econometric Reviews (2017, 36, numbers 6-9)[7] dedicated in his honor.

This interview is our acknowledgment of Essie Maasoumi's longstanding contributions to the econometrics profession as an influential researcher, a dedicated teacher and mentor, and a scholar with deep and wide perspectives on economics and econometrics. He has been an editor of Econometric Reviews for three decades and transformed it from only a review journal into a regular journal, considered one of the top five core journals in econometrics. Also, he is on the Board of Journal of Economic Studies and the Advisory Board of the Info-Metrics Institute. Essie is a Fellow of the Royal Statistical Society, a Fellow of the American Statistical Association, a Fellow of the Journal of Econometrics, a Fellow of Econometric Reviews, Founding Fellow of IAAE, A Senior Fellow of Society for Economic Measurement, among others, and appears prominently in various hall of fame rankings in econometrics. In addition to his academic talents, Essie is a great communicator and discussant on a wide variety of socio political and philosophical topics, and a very gracious host!

Could you tell us about your background and personal-education history, teachers, favorite subjects, goals, contemporaries?

I obtained an HS diploma in Mathematics in Iran, and then did some General Certificate of Education (GCE) courses in England, before admission to the LSE in 1969 to do a B.Sc. in Mathematical Economics and Econometrics, a three-year undergraduate degree. In 1972-73 I did a one-year M.Sc. degree in Statistics, also at the LSE. Then I was admitted by Denis Sargan and the LSE to do a Ph.D. in Econometrics, 1973-77. At all points I applied to a bunch of other schools just in case LSE would not take me! I was keen to stay in London, until it was no longer economically feasible. I had the enormous benefit of a very large scholarship, and from 1975 I was part time faculty and lecturer at the LSE and later a lecturer at Birmingham. Then I became very poor immediately upon graduating since I had to pay back the big scholarship. I was determined

[7] Phillips, P. C. B., & Ullah, A. (2017). Econometric Reviews honors Esfandiar Maasoumi. *Econometric Reviews*, *36*(6-9), 563-567.

to do research work and be a teacher. In the HS I had diverse interests, but specially in mathematics and sciences. I also loved history and geography. I attended a very exclusive HS, Hadaf, and somehow managed the tuition with competitive grants and such. My mother was a single parent and a poorly paid teacher. I thought I would be an engineer, and won national competitive entrance exams in those areas. But the Central Bank competitive exams and scholarship to go to London to do Economics/Accounting won out. Bahram Pesaran, Hashem's younger brother and I were school mates at the LSE for all our degrees.

What are the main themes in your work, earlier on and more recently?

In Econometrics, themes began to emerge in my Ph.D. thesis. These included: distribution of estimators, predictors, etc. My advisor Denis Sargan, and other faculty at the LSE, were keen on precision and rigor. One popular topic was exact sampling distribution and inference, especially in simultaneous equations models (SEMs). So I began with that and moved to reduced forms of SEMs and forecasting. Misspecified models, model uncertainty-identification-causality were other major themes. Also, improved estimation, mixed estimators and predictors, and Stein and ridge-type estimators, to deal with model uncertainty. Some of these are more recently taken up in "data sciences" and "big data", machine learning! I would include model averaging to which I have returned in recent years. Denis Sargan, Ted Anderson, Peter Phillips, Henri (Hans) Theil, Jim Durbin, Arnold Zellner and Bob Basmann were major early influences in econometrics, and Amartya Sen and Arrow in economics. A privileged life of friendships with, and generosity from many of them, followed.

Soon thereafter, I began to delve into welfare and well-being, multi-dimensionality, inequality, poverty and mobility. Amartya Sen, Tony Atkinson, Tony Shorrocks, Stiglitz and Gorman influences emerged over time. Still the main theme is distribution of things, outcomes, and how to assess them. I recognized early the need for measures like entropy, and the need for metrics for ranking and contrasting distributions. This led to learning about information theory, entropies, and stochastic dominance rankings and tests, later followed by treatment effects, program evaluation, and the potential outcome paradigm. The focus on distribution of outcomes rather than "average effects", inevitably led to distribution metrics like entropies and stochastic dominance techniques. I have always adopted a decision theoretic approach to program/policy evaluation. This was inspired by my interest in welfare theory and inequality. These themes led to an extensive program

of training students with my colleagues (e.g., Dan Millimet) at SMU, rather quietly at first when most of treatment effect literature was more focused on "average treatment effects".

Nonlinearities and their relative neglect were also a natural line of thinking when one is concerned with heterogeneity (whole distributions) rather than conditional mean, linear regressions and models. It is also a fundamental issue of mis-specification, of first order, a central theme. Nonparametric examination of dependence and other properties of possibly nonlinear processes led to a series of papers and work with Jeff Racine, Clive Granger, and a community of students and colleagues. Information theory concepts have been central to examinations of generic dependence, symmetry and reversibility.

My most recent works include gender and racial gap, multidimensional poverty and well- being, mobility (social and intergenerational), inference when all models are misspecified, and machine learning.

A challenging question is: what does it mean to do model averaging when all models are globally misspecified, and what is a "partial effect" in such a situation? The themes are virtually unchanged since my thesis work. Aggregation, be it of models (for averaging!), or for multidimensional measures of well-being, entails "information" aggregation. This too seems ideally done with reference to whole distributions and information concepts that are natural tools for it.

These themes appear in your early publications which indicated your interest in studying whole probability distributions of economic variables instead of just moments, and that you had convictions that all models are misspecified. What were the basis of such initial fundamental thoughts?

Thank you, Aman. In my early years structural models meant linear SEMs, with their Reduced Forms (RFs) as forecasting media. They still are! So, I decided to look at the RFs. Initially I was motivated by technical issues like infinite finite sample moments of RF estimators and, naturally, forecasts that are based on them. I then realized the important relationship with the "structural" model was a complex and pivotal issue. How do economic theories and a priori models inform RFs? In the early to mid-70s these reduced form estimators were being looked at primarily by Denis Sargan. Michael McCarty would be perhaps the only other person, as I learned later. I immediately saw that there was a problem in terms of pinning down the many uncertain structural

models that relate to the same RF. I began to speculate that maybe this is the reason for poor forecasting that was and is endemic. In those days the technical notion of poor forecasting to me was infinite moments. Sargan had just established that a lot of these RF estimators had infinite moments. So forecasts will have infinite confidence intervals, even for seemingly correctly specified models! For me the ideas of finding methods that don't have bad statistical properties and model misspecification and uncertainty were intertwined. Unrestricted, simple LS estimators of RF don't have infinite moment problems. It's the inversion of the a priori restricted IV type estimators into RF that have a problem. This is beautifully exposited in the recent paper by Peter Phillips (2017, ER)[8].

My solutions in terms of mixed RF estimators weren't so much about mixing estimators but about mixing information from dubious structural models. This motivated the question of uncertainty of models, and the notion that one is dealing with a set of misspecified structural models, with the same reduced form. I have recently returned to this same paradigm with more sophisticated tools and concepts in joint work with Nikolay Gospodinov[9] on model aggregation and averaging, when all models are "globally misspecified". This is in contrast to important work on "model ambiguity" that is being done by Lars Hansen, Marcet, Bonhomme, Tom Sargent and a number of very good young people. In the latter, "local misspecification" is adopted within a small order of magnitude (in sample size) of a single "reference model" that shall be used for inference.

To begin with I had no idea that, in fact, bounding the degree of model uncertainty held the basis for various mixtures solutions. I was trying to determine the degree of certainty in a priori models, based on statistical tests, and use that to establish inequalities that may prove useful in obtaining finite sample moments. It was stunning that I ended up with mixture methods that had also better finite sample properties, better tail behavior, and addressed model uncertainty. I should add, parenthetically, that the lack of finite sample reliability of asymptotic tests to adjudicate model choice was very much a motivating issue. The proposed mixed methods did not accept that such tests could resolve model choice questions. So there were a mix of complex modelling issues, inference issues, and finite sample vs asymptotic accuracy issues, which were to be considered

[8] Han, C., Phillips, P. C. B., & Sul, D. (2017). Lag length selection in panel autoregression. *Econometric Reviews*, *36*(1-3), 225-240.

[9] Gospodinov, N., & Maasoumi, E. (2016). Entropy Based Aggregation of Misspecified Asset Pricing Models.

simultaneously. I developed a bad habit of writing verbose papers, trying to give sensible comprehensive discussions of these issues, often unsuccessfully! The fact that economists and even statisticians were not very familiar with Stein like and other mixed estimation methods, or generally receptive to Bayesian interpretations and techniques, was an additional impediment. My first term paper took four years to get published in Econometrica (1978). And that was the easiest run for any of the other papers on this topic.

So this is how you worked on the Stein-type combined estimator (Econometrica 1978), and ridge-type shrinkage estimators (1980, JoE), and Generic Reduced Form (GRF) Estimators (1986, JoE)[10]. Were all these works the basis of your convictions regarding misspecified models and how to combine them?

Yes. I think, you know, in terms of the restricted and unrestricted models in the simple regression case, mixture estimators were being proposed. So you're exactly right, going back to the 50s, Charlie Stein had looked at this issue of the properties of the unrestricted least squares. And how certain forms of prior restrictions for testing could possibly be used to decide about the mixture, with the classical unrestricted estimator of the multiple means. I came upon that work later and used it as a justification. I had no idea about that literature at the time. So I was thinking of mixtures and pre-testing, to use inequalities like I had seen in finite sample distribution literature and one of Denis' papers. In Stein and related work all the combined estimators have finite moments, including in the normal linear regression model. It should be said that Stein and other mixture estimation procedures are principally focused on improving estimation efficiency. They are not concerned with misspecification and incorrect priors. Traditionally, mixed estimation is concerned with how to produce estimators with lower quadratic risk. I had proposed a mixed (combined) estimator and proven it has finite sample moments even though some of the component estimators were known to have infinite moments. I had a heck of a time with Econometrica referees and at seminars on this issue! How can this be? The reason is that the mixing weights are not constants. One way to reason is this: The tests of hypotheses like Wald's are ellipsoids. Once you report a model that has not been rejected, the whole ellipsoid (normalized quadratic Wald test) is less than some kind of constant, the critical level of the test! This bounds the moments of the mixed

[10] Maasoumi, E. (1986). Reduced form estimation and prediction from uncertain structural models: A generic approach. Journal of Econometrics, 31(1), 3-29.

estimator.

But I had challenges in convincing people! I began to research these pre-test estimators. And from there it was an inevitable convergence on the Stein estimator and other mixtures and ridge regression and things of that kind. I had become a “data scientist” in 1975 and didn’t know it for the next 40 years!

This 1978 paper that you cited was the very first paper I wrote in my first year as a student. And Denis had been away at Yale in my first year. I was petrified of what to show to him as progress! I had seen a paper of his about infiniteness of the moments of 2SLS and 3SLS reduced forms. There was this other paper that had always impressed me, I think it was maybe by Mariano or other students of Ted Anderson, about exploiting certain inequality bounds on functions of random variables to show certain other functions (like estimators) have finite exact sample moments. And those were the basics that came together in that 1978 paper. A curious thing about that paper, I've never presented it at a regular seminar! It was immediately submitted (in 1975?) and it was with Econometrica for three or four years. Maybe major journals read a paper from an LSE graduate student much more seriously than a faculty member at a school like USC, which was my first job in the US! My very first job was a lecturer at the LSE which could not be continued since my work permit was refused by the Home Office!

So I think my second question is then partially answered, that you've been working on other combined estimators and the problem was initiated in your thesis. In this connection, do we think that your works on the ridge type estimator, mixed estimator, Bayesian estimator, and the GRF estimator, among others, were also the basis of your convictions regarding misspecified models and how to combine them?

Very much so. They all sort of serve the idea that all models are misspecified, and what can you do when this is so. The interpretations for mixed estimators suddenly began to show themselves as a lot more helpful in terms of understanding how model uncertainty manifests itself, and how it can be incorporated and accommodated, using frequentist tests, or Bayesian interpretations. So I have favored a pretest estimator interpretation of mixed estimators and even model averaging. In more recent times, a set of models around a “reference model/law” are defined by a Kullback-Leibler (KL) “disk”. This is in works on model ambiguity. The size of such a disk could be defined by a Kullback-Leibler (likelihood ratio type) test and its significance level as its

bound/disk. Like in my 1978 Wald test[11], one can define the bound, and even let it move with the sample size, to partially reflect learning from larger samples. This will also accommodate the idea that we never reject or accept after a test with certainty. This idea of letting the critical level move with the sample size was suggested to me by Peter Phillips as early as the mid-70s in his private and extensive comments on my paper. There are very closely related old Bayes' estimators which are mixtures of the frequentist estimator and the prior mean. Sort of the posterior mean as a minimum risk mixture estimator. So, most of my papers that were written in that period and for some time after, all became different ways of trying to handle the same problems either by Bayesian formulation of the uncertainties (as in the GRF papers) or by pre-testing. The latter is frequentist, a way of addressing or reporting degrees of uncertainty, how badly a model is rejected and using that as a way of weighting mixtures. By the way, this fills a gap in the original Bayesian paradigm which did not evolve to allow prior weights on competing laws/models! Hierarchical models and mixture laws provide a good paradigm for that. There is some ambiguity about how one would conceive of a priori information about parameters of misspecified models!

The Generic Reduced Form estimators/models that you cited reflect my Bayesian thinking on formulating an ambiguous relationship between many structural models and a reduced form. And I worked on that and not particularly successfully. I made the transformation between structural coefficients and reduced forms a random relation. Lots of structural models satisfying the same observed reduced form (observationally equivalent models) with an error, with a finite variance. This is like the KL disks that define a set of possibly misspecified models, around a "reference" model, as in the recent literature on model ambiguity. I called the single observational equivalent the GRF Model. It is actually a useful formulation for interpreting the so-called "Bayesian VARs". People who would later write on the latter methods on VAR were not impressed by my GRF methods in the early days. These models also produced mixed estimators, rather like Bayes-like decisions, and estimators that are proposed in Bonhomme and Weidner (2018, unpublished)[12]. Unlike the Stein-like reduced form estimators, GRF had no testing, just typical posterior type mixtures. At about the same time Ed Leamer had begun to talk about the uncertainty of models and specification searches. And he was thinking in terms of how different models could be weighed

[11] Maasoumi, E. (1978). A modified Stein-like estimator for the reduced form coefficients of simultaneous equations. *Econometrica: Journal of the Econometric Society*, 695-703.

[12] Bonhomme, S., & Weidner, M. (2018). Minimizing sensitivity to model misspecification. *arXiv preprint arXiv:1807.02161*.

by their prior variances, as it were. This was superb work and very helpful in sorting out issues. I think at some point he called it "honest Bayesian" methods. And he was also kind of coming out with mixtures as an interpretation of dealing with uncertainty. I was very challenged in publishing those parts of my Ph.D. thesis, especially the GRF estimators. Several rounds of rejections from Econometrica, and several years with Journal of Econometrics, until I referred to my estimators as being "à la Leamer". They got published in JoE, one as a lead article, but seemingly did not have much impact. They were certainly not cited much. One of my editors, Chris Sims (coeditor, Econometrica) disagreed with his (apparently Bayesian) referees on the GRF method. He opined that the method "contradicted the fundamental logic of simultaneous equations models". In those days Econometrica coeditors, Sims and Wallis, had published a policy on appeals: Go to another journal, another editorial crew.

But at that time when you wrote this 1986 paper, you already had your PhD.

Yes, I was done by 1977, and moved to USC. The GRF related papers were published in JoE in 1986 and later. One was called "How to Live with Misspecification If you Must?"[13]. It contains some brilliant examples provided by Peter Phillips that show misspecification situations that are consistent with the GRF formulation of observationally equivalent misspecified models. He should have been a coauthor. In fact, the idea of calling these types of reduced forms "generic reduced form" estimators was suggested by Peter. Appearance of "random coefficients" or varying coefficients can be seen to be due to misspecification and nonlinearities! I interpret "regime change" models that are now so popular in macro in this way.

Denis Sargan, if I recall correctly, in that period was more involved in existence of moment issues, for the 2SLS, 3SLS, FIML estimators, and validity of Edgeworth and Nagar approximations.

Yes. He had several papers on these topics, some published in Econometrica in the 70s, and some in his collected writings which I edited for Cambridge University Press, in the mid-80s.

How did he take all these topics you were working on, and how much interaction time did you have with him?

[13] Maasoumi, E. (1990). How to live with misspecification if you must. Journal of Econometrics, 44(1-2), 67-86.

He was exceedingly generous with his time for me and others and seemed to very much like what I was doing. He helped a lot and often. His focus was the approximate finite sample distribution theory and validity of approximations. He was at the time also beginning to work on the exact finite sample distribution of IV estimators generally speaking. Robert Basmann had done superlative work on special cases for 2SLS. It would be Peter Phillips who provided the complete solution to exact IV distribution in the general case of SEMs. But as part of that program, you are right to say Denis was proving IV based estimators had finite sample moments only up to certain orders, depending on degrees of over-identification. The reduced forms estimators didn't have finite moments, except the full information MLE case. So his focus was very much mostly on the moments of these RF distributions. And one of the main topics in those days in finite sample theory was Nagar's moment approximations. Indeed, how I had come to write my thesis under Denis was intimately connected to Nagar's work and existence of finite moments. For my master's thesis, I was simulating Nagar expanded moments for 2SLS with different degrees of identification. Some of my experiments were cases in which 2SLS is known not to have any moments. I had read that in a paper by Basmann and decided to do my Master thesis on it, under Ken Wallis. You'll always get finite numbers in simulations, of course. Because it was Monte Carlo, we could see that the Nagar expansions were doing well! In cases where the moment is infinite, the simulated moments behave very badly. It is also a case where asymptotic theory misleads you. This is how Denis kind of took me on as a student looking at these simulation results. It's where he began to think of the meaning of these finite sample expansions, and role of resampling (simulation and bootstrap) approximations, including for estimators that had infinite moments. Denis's paper[14] (Sargan, 1982, Advances in Econometrics) (reflected in an appendix in my 1977 thesis) is the first observation about a phenomenon that Peter Hall exposited for the bootstrap and Edgeworth approximations. There is a sense in which these resampling methods are in fact estimating higher order approximations. And to this day I debate this result with people, such as Joel Horowitz, even though one can't quite rigorously show it in many cases. The reason these resampling methods tend to do better is that they are capturing these higher order approximations. We could see in my master's thesis that the simulated moments were almost invariably closer to Nagar's expansions. While there is a proof, for bootstrap, that higher order refinements are obtained for "pivotal"

[14] Sargan, J. D. (1982). On Monte Carlo estimates of moments that are infinite. *Advances in Econometrics*, *1*(5), 267-299.

statistics, it has not been shown to be untrue for non-pivotal statistics, as far as I know. Of course these are all standard cases, and it is known that the bootstrap may not do well in some non-standard cases where subsampling and other methods can be used.

When you were thinking about studying distributions, did you think at that time about specifying conditional distributions? I think models are really reflections of the characteristics of the conditioned distribution. For example, the conditional mean is a regression model. You were thinking misspecification issues in general, or misspecification of the conditional mean (regression)?

No, initially the importance of conditional distributions wasn't reflected in my work. Specification analysis was all too focused on what happens to the conditional mean. More rigorous discussion of misspecification is fundamentally about not looking at the right joint distribution and its conditional distributions. Theil's specification analysis in a lot of ways has been presented mechanically. Given a conditional mean, a vaguely defined question is, what if another variable should be in it? One way of thinking about this would be that there is no such thing as a misspecification here. One decides which joint distribution and which conditional distributions to examine. The question of "Which variables" should be decided by the analyst. In fact there were papers by Aris Spanos and Ed Leamer showing that the Theil analysis of omitted variables was/ is a little abused. It's still a meaningful counterfactual analysis. What if other variables impinge on the observed outcome? This means that conditional means of different distributions are being contrasted, and the properties of estimators of one conditional mean parameter are examined for a parameter in another conditional mean. If I choose a conditional distribution of one variable, y, conditioned on two x's, say, x1, and x2, and employ the data to study corresponding conditional mean, there's no misspecification. It's what I claim about the results, substantively, that could be challenged as empirically inadequate. One can do conditional regressions quite robustly, say by nonparametric (NP) methods. You are one of the leading scholars in NP. I can get the conditional distribution of a y variable given the two x's, and there's no ambiguity about it, and I can get its mean, the quantiles, and so on. So there is really little room for discussion on variable choice in that frame. It is my chosen joint distribution. It's a problem of "observational" data and the underlying "experiment" that generated the data. It may be that one has considered the wrong joint distribution.

Your work in Econometrica (1978) has motivated Hansen (2017)[15] *to develop a combined estimator of OLS and 2SLS estimators, and Peter Phillips (2017) has emphasized the significance of your contribution. Both of these papers are in the special issue of Econometric Reviews in your honor. Do you have any comments on them in relation to your responses to the above questions?*

So first to Peter's paper. It's a very significant revisit of this whole issue of forecasting from structural models. He goes beyond and is so much broader than the 1978 paper and the work we just discussed that I did. And his technical approach to it is also different and much more rigorous. In some ways, it's true or more faithful to the notion of going after the whole distribution of things, rather than the moments, which I had focused on in the 78 paper. What Peter does is to consider the issue of forecasters, distribution of forecasts, their finite sample distribution, and how that associates with the asymptotic distribution as an approximation.

He then deals with the special asymptotic of it in the case of many moments or identification or weak exogeneity issues, and weak moment conditions. He uses that setup, which he had used in the 80s, in another context to really expose what it is that goes wrong with the statistical properties of forecasts, even though he uses the context of my '78 paper in simultaneous equations. Any time you have a ratio, like the final form of dynamic models, you have the same issues. And he looks at that kind of ratio as having a singularity with a nonzero probability, and how that impinges on the distribution of forecasts. And, it becomes a clever observation in his context, how it is that mixture estimators such as mine can impede that passage through infinity, and deal with lack of finite moments. It is brilliant. He shows how it is that you bound away to exclude the singularity from the set of possibilities, those points in the space that are causing the Cauchy-like behavior. And in his paper, that comes out so beautifully, so eloquently. And fundamentally, his results are asymptotic. But like his papers on under identification, the asymptotic is derived from the finite sample distribution. Another level of approximating the same thing. It is all in the framework of structural models, and it's so beautifully revealing because that translation/transformation from the structural to reduced form is where we also discuss identification. And it's the weak identification in those settings that causes a problem. It's one way of looking at the uncertainty of things and how information from a structure can be brought in to various degrees. So, in fact, that's what mixture estimators are doing. In that context. Phillips

[15] Hansen, B. E. (2017). Stein-like 2SLS estimator. *Econometric Reviews*, *36*(6-9), 840-852.

(2017) rightly points to more modern terminology and approaches like LASSO. And LASSO now even uses or can use modifications that use economic theory or a priori information about models to affect how you penalize. So modern "big data" is included in this. You can have many variables, or many moment conditions, and many weak conditions with many weak exogeneity assumptions or relationships. And all of those through the method of LASSO, which has a mixture identification itself as well, exposes why these mixtures worked in my setting. And would work in a lot of other settings.

Bruce Hansen's paper is more of a modern, direct application of the Maasoumi (1978) paper on pre-testing. In my paper I used tests, letting them decide what kind of weighting you give to prior information/restrictions, and how that bounds the mixed estimators so they have moments. Bruce is focused on the structural coefficients, not forecasting (reduced forms). And he's concerned with asymptotic results. He also uses asymptotic results for weak moments, and exogeneity settings. But he's very generous and kind to attribute the idea to me. Bruce combines structural coefficient OLS and 2SLS on the basis of the outcome of a Hausman test of exogeneity, the Durbin-Hausman-Wu test, which is an example of the Wald tests I used in my 1978 paper. In the latter, the Wald test was re-cast in terms of reduced form coefficients. The Hausman test is directly about the structural model, and so can be used to decide what weights to use. And these weights depend on sample value of the test and the significance level. It's actually very Bayesian in itself as it reflects degrees of belief in models. Model uncertainty is rarely accounted for in our field, and mixed estimators can be seen as a way of doing that. And then Bruce finds something very interesting, similar to some other papers he's had. He proves an asymptotic version of Stein's classic result. Stein had this fantastic puzzling result that maximum likelihood (least squares) is dominated in mean squared error if you have more than two regressors. And one gives up something on the bias of course, that's your cost. Bruce proves the same thing asymptotically that the mixture, the Stein-like mixture that he defines, is superior in terms of a certain asymptotic weighted mean squared risk, to 2SLS, which would be the efficient IV. Stein's was all in the context of normal means and a linear regression model, and finite sample distribution theory. So it's a pretty significant result and extension of Stein's.

But the issue of inference with these estimators is still a problem.

You are so right to point to it and appropriately so. So this idea of dominance is still focused on moments and risk. In particular, quadratic loss and first and second moments. The question of what to do with issues of inference gets back to the more pressing problem, what is the distribution of these decisions, right? One response to that is, there isn't very much done. The exact distributions of these pretest estimators, mixture estimators, shrinkage estimators, are extremely complex. On the more optimistic side, there is the device that I used in my work on ridge regression, and in the '78 paper, to set things up in a way that there would be an asymptotic convergence of the mixture distribution to the distribution of the efficient estimator, or the known estimator, or something else, in a limiting sense. So my Stein estimator has the same asymptotic distribution as its efficient component, 3SLS. The ridge factor in my 3SLS ridge estimator for structural models is of order 1/T (sample size), so it converges to 3SLS. I'm not at ease with these devices right now and I'm doing things very differently. The idea is this: if the correct model is in the set of models to be discovered as your sample size increases indefinitely, you will converge on it. If your methods are designed to do that, the true DGP is included, then you should converge to it in your inference methods, whether estimation or testing. That's not something I'm happy with, and we talk about that when we discuss model averaging over "globally" misspecified models. Otherwise, when the sample size goes to infinity, we'll pick the right model. If the restricted (a priori informed) model is not correct, weights will go to the unrestricted estimators. So predictions could potentially totally ignore economic and structural theory and information.

This device of making ridge factors and other adjustments to be a diminishing order in the sample size is surely not unique to me, but I seldom see that trick used or noticed. So you can have this LASSO factor or ridge factor be anything you want including some constant divided by your sample size. In finite samples, you are doing the shrinkage, but asymptotically, this shrinkage factor vanishes. So the mixture or LASSO estimator has asymptotically the same distribution as the well-known component estimator. By the way this is what happens to the posterior mean! The prior has an order one contribution, and likelihood has order T (sample size) contribution. So, asymptotically, the role of the prior diminishes.

I should mention that George Judge and coworkers had derived the distribution of some pretest mixed estimators in several papers, and econometric textbooks. These were for ideal linear, normal, regression models with the true model assumed known.

I think your contribution in '78 is definitely very fundamental. But I was trying to trace history whereby I could know who was the first econometrician or statistician developing the concept of combining estimators? Of course we know the work of Granger and his coauthors who developed combined forecast.

There was some earlier statistical work on combined estimators. So, a lot in statistics of course goes back to Charlie Stein, Vandaele, Hoerl, and others. My 1977 Ph.D. thesis has a set of about 100 references to these early statisticians who in the 50s and 60s had worked on mixture estimators, shrinkage estimators, ridge and so on. But one paper that precedes mine had a huge impact on me in terms of thinking in a Bayesian way about these mixtures, was Zellner and Vandaele (1972)[16]. It is earlier than Zellner's Bayesian method of moment ideas. I had become familiar with the posterior mean being a mixture decision. And then at some point, I took note of this mixture interpretation in the manner of posterior means and the interpretation of why mixtures might be good Bayes decisions. But some ideas of model uncertainty were already discussed. Prior and posterior odds ratios, for instance. In econometrics Sawa and your own work in the early 70s, on k-class families in Econometrica. I wrote my 1978 paper in 1973. I was certainly not aware of Sawa's paper at the time, but became aware that he had worked on mixed estimators in SEMs, one of the few people who had worked on mixed structural estimators.

How about the work by George Judge and Bock and others on pre-testing and related estimators?

Yes, pretesting was a big deal and of concern already. George Judge and Bock were writing on it in the classical linear regression context. They were very concerned about pretesting and the impact of pretesting. And later on when I was editing a book on Denis's papers, I opened his infamous metal office closet, full of papers, some never published, and noted he had written a working paper just as a response to George Judge presenting a seminar at the LSE in which he talked very rigorously about pre-testing and what it means. We published it later in Econometric Reviews, with Mizon and Ericsson[17] and expanded on it. At the time this business of pre-testing

[16] Zellner, A., & Vandaele, W. (1972). *Bayes-Stein estimators for k-means, regression and simultaneous equation models*. HGB Alexander Research Foundation.

[17] Ericsson, N. R., Maasoumi, E., & Mizon, G. E. (2001). A Retrospective on J. Denis Sargan and His Contributions to Econometrics. A Retrospective on J. Denis Sargan and His Contributions to Econometrics (March 2001). FRB International Finance Discussion Paper, (700).

was recognized as having a problem, because every estimate ever reported in empirical settings is a pre-test estimator. Models are tested and then reported! The fact is that the reported standard estimates do not have the properties of the "apparent" estimators. This is still poorly reported and assimilated. George Judge and coauthors were doing a good job exposing this. The actual distribution of pretest and mixed estimators is pretty nasty looking and makes for difficult exact distribution inference. I was aware of George's work, but again, it was not in simultaneous equations. Bruce, in his 2017 paper, suggested I might have been the first one to work on these things in simultaneous equations. But I think Sawa was earlier, at least for structural equations mixtures, not reduced form equations and forecasting. Granger and colleagues had looked at mixed forecasts and puzzled why they worked so well, but again no technical and decision theoretic explanations. The Sargan paper sheds light on some inference issues that deserve more careful consideration in big data variable selection procedures. While penalization may not formally employ pretests, it can be given such a statistical interpretation. That opens a whole can of worms about statistical properties of such methods!

I see you have again gotten back, after having deviated for some time to other areas of your work we get back to later, to the area of shrinkage estimation with issues related to selecting many instruments and moment conditions, especially with Caner and Riquelm (2017). Could you please let us know about the directions of these developments done after a gap of 30 years or so, and how they are related to model misspecifications and averaging?

Yes, yes, so the mixture estimation stayed out of my focus for several years but not misspecification and the notion that all models are misspecified. Just parenthetically I should mention an important development with Peter Phillips. In 1982 we published something about the distribution of instrumental variable estimators[18]. When models are misspecified, we discovered a very disturbing result that in some cases the asymptotic distribution does not exist. Later Alastair Hall and Inoue (2003)[19] extended this result to the GMM. This was a very general result, it has also to do with too many moment conditions identifications.

[18] Maasoumi, E., & Phillips, P. C. B. (1982). On the behavior of inconsistent instrumental variable estimators. Journal of Econometrics, 19(2-3), 183-201.

[19] Hall, A. R., & Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, *114*(2), 361-394.

The work with Caner and Riquelm kind of revisits that. It is constructive about situations with many moments, a big data setting, LASSO type setting. The question is how do you penalize to pick a subset, or a penalized set, and how do you do that in the context of misspecified, inexact moments, possibly under-identified models and for GMM estimation. So we couldn't get all of those problems in and prove theoretical results. So that paper is all simulation and the theme of misspecification is maintained, but it requires very careful thinking especially in the context of "big data" and machine learning. One needs to think through the objective since estimation of partial effects poses very different, albeit related, challenges compared with accurate outcome predictions.

In a series of papers with Y.P. Gupta in the early 80s, we looked at what happens with the misspecifications of various kinds in dynamic regression models. One result was a very simple application of Theil's omitted variable analysis, but on missing trends polynomials. And I think it's never really attracted attention, maybe it's too obvious but it's so important for de-trending. So when you de-trend data with a common linear trend, and the real trend is nonlinear, the bias goes to infinity and becomes worse with sample size! It's an Economics Letters piece and it's an inconvenient negative result about using (mis) de-trended data.

The work with Caner and Riquelm is in the GMM context, which is also the context in the current work with Nikolay Gospodinov on model uncertainty when all models are misspecified. I am struggling with what does it really mean if the true data generation process is not in the set of averaged models? And the more I read, the more I realize that, either directly or indirectly, most people include the true DGP in model selection procedures, and Lasso and all the others, at least in a limiting sense. Bayesian model averaging ordinarily includes the true model, and one converges to the "true DGP" asymptotically. The "true DGP" has to be there for convergence to it. It is challenging when the true model is not included. What are the meaningful objects of interest and inference? As we know, even in the discussion of omitted variable linear regression, there's a bit of confusion because of notation. We use the same notation for the beta coefficients (partial effects) in both the misspecified model and the correct model. That's already misleading. And it causes a lot of confusion. Once you use different coefficients you realize there are issues of defining the object of inference. Which partial effect, which conditional distribution and corresponding model restrictions are the object of inference?

In contrast, model aggregation focuses on an oracle (optimal) aggregator that is itself the legitimate object of inference. It is similar in nature to the so called pseudo-true parameters. In a parametric setting we define these pseudo true values as sort of optimizing members of the model class. Every model and every estimation criterion, however, defines different pseudo true values! This is a conundrum when you have model uncertainty. Model selection is seen as suboptimal aggregation, and somewhat confused as to objects of inference, especially as it concerns partial effects for policy analysis.

In model averaging we have an ideal average over a given set of models, relative to a risk function that reflects costs of forecast errors. We have to work harder and think deeper about objects of inference across competing models. These objects may, by the way, not be very interesting [LAUGH].

I have recently done some work on mixed quantile estimators with Zhijie Xiao. One idea there is to explore better forecasting, based on conditional means, but with weights coming from conditional quantiles. There are several ways of mixing conditional quantiles for improved forecasting.

Let us talk about LASSO and related selection procedures. Is it appropriate to do econometrics by dropping variables having low numerical values ignoring their economic significance? What about the bias produced from LASSO selection?

A deep set of questions. I think LASSO and similar penalization/regularization methods are generally focused on prediction. I alluded to the difficulties in statistical interpretations of these algorithms. Can one discover mechanisms and treatment effects? Perhaps, such as with "double machine learning". This is similar to Frisch-Waugh partial regressions that we show our students as equivalent to full multiple regressions. In more general models, Neyman orthogonalization provides the same partialling out step, but both the target explanatory variable and outcome variable have to be partialled out (by machine learning on big data). There is limited experience with these methods owing to the limited examples of big data availability in economics. I am curious about the practical properties of the "residual series" obtained from the two machine learning steps.

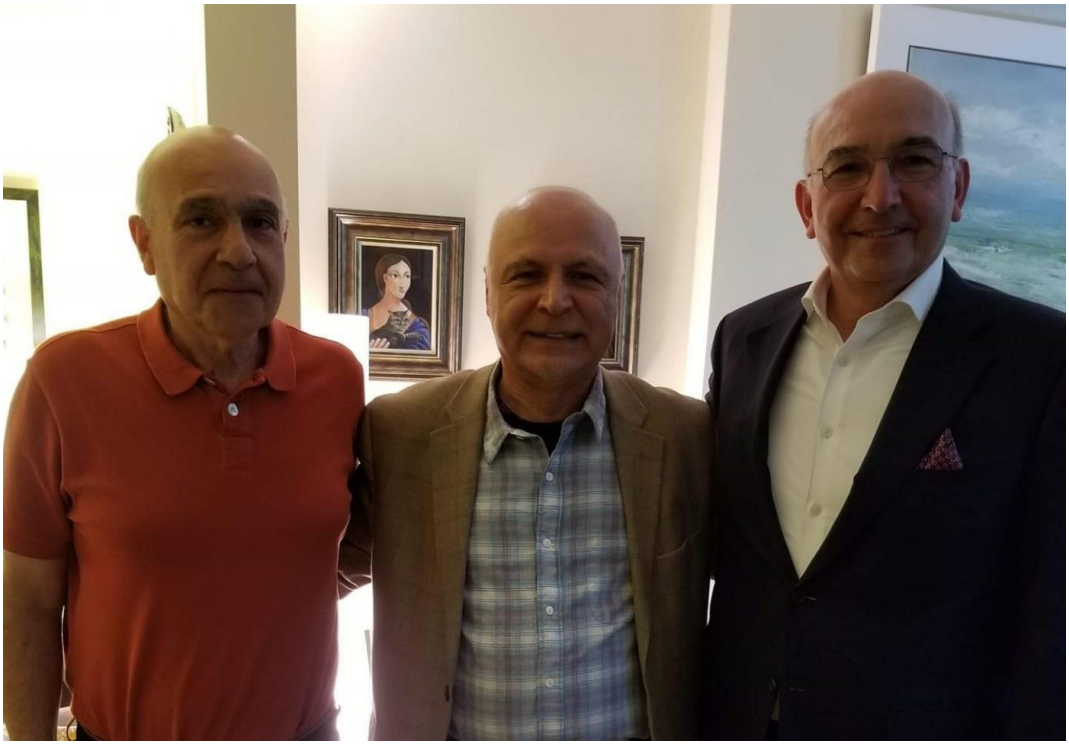
Victor Chernozhukov, Hansen, Newey, Athey and their coworkers provide state of the art accounts of these methods in econometrics.

What about the bias issue?

On the bias, it kind of becomes part of the risk considerations like MSE. The aspiration to discover *ceteris paribus*, partial effects is just that, an aspiration. In most "big data" methods one sacrifices a lot for good prediction, fitting. So the fact that each coefficient is biased is accepted and raises very fundamentally what are they biased about? Partial effects? It is not likely that we would ever get unbiased estimates of these objects. It's very difficult to now even talk about the "pseudo true" objects for the partial effects. So unbiasedness is really a complex aspiration. Here I should give a shout out to nonparametric methods and the derivative effects they are able to produce in lower dimensional situations. Semi-nonparametric and series methods are, again, good at "fitting", but not so clear in terms of partial effects. They have other issues, of course, but I don't want to sound like I am an expert on these methods. Policy decisions that require quantifications of (treatment) effects are quite vulnerable to biased estimation. Think of the debates about the magnitude of intergenerational mobility, for example. These "coefficients", partial effects, matter in discussions and comparative analyses!

Some argue that more instruments can lead to more inefficiency. Where do you stand on this? Given notable recent developments we would like to hear more from you on LASSO, shrinkage, and misspecification issues.

The idea of too many instruments is quite interesting. It's interesting you raise this because it goes back to our discussions of exact finite sample distribution theory. So this little cited paper by Ted Anderson and Morimune shows that if the number of instruments increases indefinitely, even though you have more and more degrees of overidentification, you get very poor finite sample distributions. The all important "concentration matrices" become very poorly behaved. This is sometimes discussed from the viewpoint of weak instruments. You can look at them from the point of view of what's happening to the rank of the "concentration matrix" that is fundamental to the behavior of the exact distribution. Many instruments, even though they offer greater identification degrees, can lead to very poor finite sample distributions. The weak instruments asymptotic don't ensure good small sample properties for estimators when you have too many instruments. It's very



Bahram, Essie, Hashem

insightful of you to mention in the context of LASSO, because it's again many variables and implied IVs, many of which are expected to have close to zero influence. Sparsity conditions imply that. So, I think machine learning and LASSO techniques are a lot more suited to “prediction” when mechanism discovery and partial effects are not the focus.

Your question takes us back to the origins of IV, especially as formulated by Sargan. There are two properties required of an IV. “Orthogonality” is very familiar, but the second requirement of predictive value is widely neglected, or was neglected for many years. The fixation on finding orthogonal variables leads to sometimes ridiculously unrelated variables. Of course model “fit” cannot deteriorate with even irrelevant variables. However, precision of inference on partial effects deteriorates with large instrument sets and conditioning variables. So, I have a lot of misgivings about LASSO type methods for discovery of mechanisms. They could do very well in algorithmic finance, for instance, until they don’t!

Regarding misspecification, I alluded earlier to new work by Victor Chernozhukov and colleagues. Victor considers a small set of target variables and a large set of additional controls, so like a treatment effect model, with a single treatment variable. Many additional variables may

produce better estimates of propensity scores and provide better controls. This is positive for robust specification. One can do double machine learning, like in Frisch-Waugh regressions. The idea is the same really, but one has to view estimation/optimization with respect to the desired partial effect, constrained by optimization with respect to all other loadings (like a primal of a Lagrangian optimization). This setting would seem to provide robustness to some forms of misspecification, especially since large numbers of factors and variables can be included in its big data learning steps. But the inference theory still requires the true data generation process to be included as reference, at least in the limit.

One question that I have been worrying about is this: If the partial regressions with big data methods do a good job of "purging", then the "residual" series that are obtained should be close to white noise! Right? Then I am not sure how to interpret the partial effects that are estimated between these white noise series. There is a lot of room here for spurious causality.

Relatedly, Hashem Pesaran, Chudik, and Kapetanios have a new paper (Econometrica, 2018)[20] in which they argue in favor of bivariate regressions to obtain estimates and tests on what they call "average total effects". These effects include the desired partial effect, and indirect effect of any omitted variables that may be correlated with the included variables. Notice this is the probability limit of the OLS estimated coefficient, the pseudo true parameter in Theil's specification analysis. This setup requires the true DGP to be included in the set of models, but can be robust to some kinds of misspecification. All of these works are exciting new directions in more realistic and robust econometrics, all with a more modest view of what can be inferred in misspecified models. Let me also say, candidly, that I take a rather dim view of "misspecifications" that are like "local", that is inconsequential asymptotically. This is not very meaningful as representation of "all models are misspecified". Frank Fisher (1962, Econometrica)[21] showed how this kind of asymptotically irrelevant misspecification, for 2SLS, is, well, asymptotically inconsequential! I am not sure what we learn from this. So, if some correlations are of small orders of magnitude in some sense, like as sequences in the sample size, then they can be ignored in first order approximate inferences.

Any other new developments of work on this in the big data context.

---

[20] Chudik, A., Kapetanios, G., & Pesaran, M. H. (2018). A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models. *Econometrica*, *86*(4), 1479-1512.

[21] Fisher, W. D. (1962). Optimal aggregation in multi-equation prediction models. *Econometrica (pre-1986)*, *30*(4), 744.

Yes, some new work mostly with my friend A. Habibnia, on elastic nets, machine learning with neural nets, with simultaneous L1 penalization (LASSO) and model selection (Ridge, L2) elements. We also allow an "outer loop" optimization by which one can allow data dependent selection of tuning parameters, and weights for linear and nonlinear components. This is so called Deep Learning, and we show it works really well for market return forecasting. As I mentioned before, prediction becomes the object, not partial effect discovery. I had earlier (Maasoumi 1980) suggested a ridge like estimator for large systems of simultaneous equations. These systems can suffer from near singularity both because of large size and because of other rank failures, including those coming from estimation of covariance matrices.

You have worked cleverly on aggregating sets of misspecified models using information theory (IT). This is exciting. Could you tell us about the role of information (entropy) in doing so?

Absolutely. So back to the PhD thesis days, the struggles with what topics to study in micro theory, macro, social choice, econometrics. I wondered what was the objective meaning of the "agent's information set". I couldn't find a place that addressed this nearly adequately. And so I talked to Denis Sargan. He suggested I spend time with H. Theil at Chicago. "Why don't you go there for a year?" he asked. The visit never happened because I had also wanted to get to know Arnold Zellner to learn Bayesian econometrics. But I learned that the two of them (Arnold and Hans) didn't get along. So I didn't go. Later on Hans came to USC when I was faculty there. And I mentioned that same problem to him, I remember at a lunch on a Thursday or Friday, when he asked if I had read his book on IT. I said no. He said well, why don't you read it and we'll have lunch about it on Monday. [LAUGH] And I'm still trying to finish that book. [LAUGH] It became the beginning of learning about formal IT. One interest was the direct relation between Shannon's entropy and Theil's inequality measure. The notion of flatness, uniformity of distributions. The beauty of digitization of messages, and quantifying information. So, Theil's measures of equality are from that. IT quickly began to suggest ideas of branching and aggregation. And that's really where I owe a lot to Henri Theil, because Hans' work on production and consumption was faithful to this earlier understanding of IT as aggregation. He was very big on the question of aggregation and indexing.

In those days I carried this sense of "regret" about not doing "economics" in my Ph.D. Tony Shorrocks (of LSE then and a great friend) told me early on, if you want to get into the field of

inequality, do it, but please don't produce yet another measure of inequality! So I thought, ok, I'll look into multi-variable wellbeing. Sen talked about that a good deal. And I'm a very political animal who thought about these things a lot. The idea that a dollar of income in the US gets you the same welfare as a dollar of income in Sweden, India or Iran, didn't sit well with me. With different contributions from extended family, provision of health, education, security, and all other things that are not necessarily traded (priced), it's more than a matter of real prices and purchasing power. So, I began to examine multi-variable well-being, another latent object, like well-being and happiness, and the challenge of aggregation over multiple indicators like health, income, education, access to good water, urbanization- you name it.

Fortuitously, perhaps predictably, I noticed again that one is talking about the distribution of things as the ultimate statistical information. And if I were to find the aggregate of these different components, the distribution of the aggregate had to remain faithful, close, to the distributional information of every one of the constituent components. The big thing in the space of distributions is what is "distance"? That's mathematics, right? And so divergence between distributions. That is what entropies offer! An optimal index should have the minimum information distance to its components. And that's how I ended up with the 1986 paper in Econometrica. The work I mentioned to you with Nikolay Gospodinov on aggregating misspecified models adopts a similar posture! The aggregate "model" is the aggregate of all these candidate models, as indicators of a latent, underlying Data Generation Process (DGP). The best average model is a generalized hyperbolic mean of the component probability laws.

To deviate a bit, entropy is an expected information (uncertainty) and its maximization subject to moment conditions provides distribution of a variable. But, statistically expectation (average) is not a robust procedure and thus its use may not provide robust inference.

It is like the characteristic function. It too is an expectation of a suitable function. It gives us the Laplace or Fourier transformation. In my work, especially with Jeff Racine[22], we use nonparametric density and distribution estimation to estimate entropies, which is tricky. Some old estimation methods show that the density estimator that may be inconsistent can still produce a consistent entropy estimate. But this is different from the robustness issue that you are alluding to,

[22] Maasoumi, E., & Racine, J. (2002). Entropy and predictability of stock market returns. Journal of Econometrics, 107(1-2), 291-312.

I think. It is difficult to think of why, a priori, one would have different weights for different "information" at different parts of the support, other than a probability weighted average, expected information. We are very careful with numerical integrations in our routines. They are mostly in the package NP, in the R suite of statistical software (which are freely available). If there are prior weights, one looks at "relative" entropy to those priors, but that too is an expectation.

But the characteristic function exists.

Absolutely right. So it is like the situation with the moment generating function. Entropy has that same role in that it's useful only when it exists. Such functions exist, to an approximate degree. One could operate as though they do. But the beauty of the inverse problem that you alluded to is that it focuses on discovery of exact distribution with finite entropy, the maximum entropy distribution. The inverse problem takes every bit of available information, objective from sample or otherwise, as side conditions, and gives one control over how much of it to use, efficiently. With Max Entropy (ME), all kinds of known distributions can be interpreted as being an optimal choice of an approximation, with certain side conditions. For example, for a continuous variable, if you are willing to impose information on just the mean and variance, the ME distribution is the Gaussian. Any other distribution implicitly embodies information we have not declared. I love this inverse problem. Once it is extended, it is the basis of "empirical likelihood" methods.

Since you have been an expert in entropic based information theory, I would like to ask an issue related to it. I think information theoretic econometrics based on maximum entropy is an important branch. But if you look at econometrics books and research journals its relative mention in them is very little, in fact non-existent. What are your comments on this?

You are a leading expert on IT yourself, so pressing the right buttons! [LAUGH] Because I have probably bored the heck out of a lot of audiences since the early 80s, attacking indices, averages and such. So there are disagreements about what/ how things work among people who use the same data, the same sample, and the same methodologies. This has to do with the heterogeneity of outcomes. We are used to assume that there is a representative agent, the same utility function for all, and so on. And it hasn't worked so well, in my opinion. So in the context of summarizing the ultimate scientific information, a whole distribution, we often focus on the average or median or some single index. It's incredibly convenient. It's a huge part of

communicating what we know, what we want to do. Decision makers can absorb that a lot better. And sometimes, these indices are representative, they say something useful about distributions. But often they don't. But for deviations from the average, all the things that are interesting, unusual events, crashes, poverty, mobility, averages won't do. Outcomes are different at different quantiles. This can manifest as nonlinearity, but is more. To summarize these non-average events and evolutions we need more than conditional mean regressions that dominate empirical economics and social sciences. All indices are not created equal. Entropy is like the characteristic function. It can uniquely represent a whole distribution. But entropy is a different transformation which is a lot more meaningful because it's a measure of the flatness, uniformity, risk, inequality, total dispersion, etc. Generalized entropy needs to be understood better, because there isn't just one single entropy. There are lots of inverse functions of probability as "information functions", not just the negative of the logarithm, as Shannon proposed. The inverse-Max Entropy paradigm can give discipline to many discovery problems, including the method of moments. This was shown by Kitamura and Stutzer[23], and is mentioned in my 1993 survey on IT, Maasoumi (1993, ER)[24]. And it has become really of great help to me in dealing with treatment effects in program evaluation topics, and in revisiting a lot of hypotheses in econometrics, like the general notion of "dependence", symmetry, causality and such. These hypotheses have all been popularly translated into their narrower "implications". Entropies allow us to define dependence in accordance with the full definition of independence of random variables, not by implied correlations and such. You can also test symmetry, it has something to do with the shape of the distribution, not just the third and other odd moments. I will be writing a chapter on this topic for a two-volume set that World Scientific Publishers are compiling on my published papers.

To be specific, Shannon entropy has helped to characterize univariate and multivariate distributions, but it does not seem to have ever been used to study misspecified models, except for some work on univariate distributions and use of KL divergence for testing and estimation of a given model. Should we not have more entropy-based econometrics, especially around an objective function of maximum entropy (expected information)?

---

[23] Kitamura, Y., & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *ECONOMETRICA-EVANSTON ILL-*, *65*, 861-874.

[24] Maasoumi 1, E. (1993). A compendium to information theory in economics and econometrics. Econometric Reviews, 12(2), 137-181.

KL has played a crucial role that is not widely appreciated and emphasized in teaching econometrics. I am in particular thinking of its role in defining pseudo-true values of objects, like parameters. This is very useful and fundamental when models are misspecified. The parameter value of a model which is the KL-closest to "the distribution of the data", is the pseudo-true value that is really the only firmly defined object of inference when all models are misspecified. This concept now plays a widely accepted role in at least high-brow econometrics. It is adopted in Hal White's book, for instance, on Asymptotic Theory for Econometricians and his other book on model specification. I wrote a short paper in Journal of Economic Surveys (2000) in which I quoted a result about linear projection parameters being interpreted as KL pseudo-true objects. This was earlier noted by a number of authors like Gokhale and Soofi. Unfortunately, this concept has not penetrated standard textbooks and communication of statistical and economic results. For instance, in the textbook discussion of Theil's misspecification (omitted variable bias etc.) which we discussed earlier.

David Cox and later, Hashem Pesaran proposed tests for non-nested models that are "general likelihood tests". These statistics are clearly direct functionals of the KL which can contrast between laws, nested or not.

It's quite amusing that at the beginning of econometrics, in fact, a little bit before what we consider the era of modern econometrics, the very few existing "textbooks" in econometrics had chapters on information theory. One was written by Davis, a mathematician at Indiana University. He was asked by Alfred Cowles about any research work that may put forecasting on a scientific basis. He helped Cowles to start the Cowles Commission at Chicago, involving the "correlation analysis" work of the Europeans, and "econometrics" was born. Davis wrote the first English textbook, the first modern book in econometric methods. It has four or five chapters on information theory. The inverse problem, what is entropy, what's maximum entropy and so on. Then Gerhard Tintner, a very famous econometrician, had one of the first really modern econometric texts. He had six or seven chapters on information theory. Theil had one book completely on information theory, and also in his other textbooks he would mention IT and use it for a number of substantive economic applications in production and demand systems. Later on George Judge and several coauthors like Carter Hill and Amos Golan wrote textbooks that introduced IT in a central way. But it seemingly had little impact on the practice of econometrics or economics. Theil used entropy for income inequalities, and that was durable. It's still one of the main popular measures of

inequality. The current situation is more encouraging. I think we have come a long way since your own survey and mine on uses of IT in econometrics. Sometimes IT methods are used without recognition of their pedigree. For instance the very popular divergence measure between distributions known as the Cressie-Read measures is the generalized entropy measure. Within econometrics, things like empirical likelihood, or exponential tilting, generalized empirical likelihood, are from maximum entropy empirical distribution that is more informed than the unrestricted maximum likelihood and the old-fashioned empirical CDF. So, instead of probability estimates as relative frequencies, we use these additional moment conditions, or the fact that probably has to integrate to one. You get empirical likelihood estimates, and of course you could use that vehicle to also estimate other related objects, like parameters. And this a fairly significant development in econometrics. The use of empirical likelihood is a lot more widespread now.

In finance, there is a mini explosion of interpreting and using IT measures, for optimal pricing of risk. Hansen and Jagannathan's delta[25] squared from GMM is now given an information theory interpretation as a distance within the Cressie-Read family. We go over these in the recent paper with Gospodinov on best model averages. And the stochastic discount factor models are being tested and justified on the basis of information criteria.

At least within statistics, there's work that is related to entropy to measure dependence, serial dependence, and generally. I've tried to publicize the lack of distinction, if you like, between copula and the well-known information divergence (gain) between a joint distribution and the product of its marginals. When you think of Sklar's theorem for copulas, the joint distribution is equal to the product of the marginals, embodied in a factor that picks up the joint dependence. So the ratio of the joint to the product of marginals is this factor that we call copula. And the expected value of the logarithm of that ratio is the central notion of information distance. It is in fact the entropy of the copula density! This entropy of copula density is exactly what we use all the time, you and I, to measure statistical dependence in IT! This is not a new concept at all, except for parametric formulations of copulas that are very helpful. The parametric copulas are big, right? It's so convenient, you can do maximum likelihood estimation with them. You can see why we have the non-parametric version of copulas. In IT we estimate joint distributions non-parametrically, and IT measures are accordingly more robust. And so we already used, and continue to use non-

[25] Hansen, L. P., & Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy*, *99*(2), 225-262.

parametric copulas, before copulas [LAUGH]. The package NP in R does all of these. It was expertly developed through great work of my coauthor and your former student, Jeffrey Racine.

So the answer to your question is, yes, it's disappointing that IT is not more widespread. The more subtle answer is there is a lot more of it than meets the eye. The other part can be characterized as the fault of not educating. We don't teach it, they don't know it, so it won't show up in research. It used to be in the textbooks and it impacted a lot of researchers at the time. And then it wasn't in the textbooks, certainly it wasn't in the programs, and so people don't know about it. Because it sounds really abstract and people come to it through mathematical writings. They come to this notion of entropy in quantum physics, and it must be very difficult math. And so that's a barrier, but textbooks can deal with that by demystifying. There is a specialized journal, Entropy, which is less familiar to economists.

Processing speed is the real revolution, more than algorithms and techniques. Combined with digitization and information theory concepts for minimum length messaging and encryption, this has transformed our lives. It's how fast you can process these things that has made a difference. "Big data" techniques, as an example of computational methods, are all feasible because of digitization, encryption, and information methods. In econometrics, it deserves greater emphasis in our textbooks and courses.

You think it's a problem of teaching?

You and I teach it all the time. George Judge and Amos Golan teach it. In my classes we find a place for it. It is so evident when you give seminars. There is this barrier with lack of knowledge of IT, you have to sort of cross this bridge before you can get to what you want to actually talk about. We have to find more ways of showing it as capable of solving substantive economic questions.

This is also related to one of the themes that you and I talk about: how the space of distribution is often a better space to work with. It's so much more accommodative of the notion of heterogeneity. In much of economics we assume massive degrees of homogeneity. Agents have the same utility, they all optimize equally well, and everyone is equally informed in the same way. Now more people admit that's not really working very well, so they have begun talking about multiple equilibria arising from multiple information sets, may be processing information differently. When you sell people the idea that they have to deal with heterogeneity, then they

have to face the challenge of examining and comparing distributions of outcomes. And once people realize the need to look at the whole distribution of outcomes, it may be seen that nothing delivers like information theory in terms of distribution metrics.

Have you thought about writing a book on this?

Now based on examples of uses of IT that we have been demonstrating in the field, it would probably make writing such a book more fun. I have two books under contract. They are supposed to be collections of my papers, in addition to two introductory chapters, and possibly some new chapters. I will need a real sabbatical, from everything, to accomplish this goal. I find writing books very daunting. Also, my theory is that it takes great patience to stay with the same topics and focus. I find respite from one topic to come from a totally unrelated topic. This is how I rest!

Now let us move on to your influential work on multidimensional poverty, inequality, mobility, and well-being. Tell us about the path breaking paper in Econometrica (1986): its origin after your publications on shrinkage estimation we have discussed above. It looks like your work on multidimensional poverty is originated in your original thrust on studying distributions. Is this correct?

Yes, it's back to this theme of working on the distribution of things, and multiple indicators for a latent object like well-being, or the "Data Generation Process". Topics of mobility, poverty traps, inequality, generally, have become of universal concern. Their topicality can be measured in many ways, including the number of papers published in the top 5 economics journals, faculty hires at major universities, the number of university and government centers and institutes that have been established to study these important global issues.

The 1986 Econometrica paper proposes the ideal generalized entropic aggregate of multiple indicators of well-being. Each indicator, like income, health, education..., has a distribution as its full contribution, statistically speaking. So, it makes sense to have an aggregator whose distribution is closest to the distributions of its components. It is the index number problem with an IT solution. By construction, all other indices/aggregators must be suboptimal. Based on this aggregator, I proposed a measure of multidimensional inequality that benefitted from many decades of work on axiomatic properties of ideal inequality measures.

Please tell us about your work on stochastic dominance, its origin, and the publication of your highly cited paper in Review of Economic Studies (2005). Did you feel happy about this publication since it is related to studying distributions and their ranking?

Technically speaking, one way to compare distribution of outcomes like income, wealth, health, education, etc., is by scalar metrics such as means, medians, the Gini coefficient and entropies. Those are cardinal measures which are equivalent to picking specific utility functions or decision functions. For instance, every entropy connects to a decision function. So every entropy is very subjective, as are means and medians. How can one get away from that and do uniform ranking, uniform comparisons over entire sets of preference functions? Ranking outcomes and decisions is inevitable in science. It's part of optimization. Every time you prefer (compare) something over another, one policy outcome over another. How can you resolve disagreement over implicit, subjective utility (decision) functions? Stochastic dominance presents itself as an opportunity, as a potential. Can one rank two *distributed outcomes* no matter what the utility function? Initially, my concern was with income distributions, and tax policy outcomes. Dominance rankings had been done graphically for a long time in finance. The ideas of mean variance analysis go back to the early 50s. In fact Arrow is probably one of the first to show that (quadratic) mean variance analysis is axiomatically flawed. It's not compatible with the axioms that he had used to produce existence or non-existence of equilibrium. For many scientists it's difficult to part with mean-variance comparisons. It seems like 99% of science and economics is based on variances [LAUGH]! And we have optimum portfolios, estimation decisions and so on, and even quadratic testing, standardized quadratics like Wald's. Comparing prospects is also central in finance, based on Gaussian distributions and/or quadratic risk functions, neither of which is appropriate to patently non-Gaussian returns, and non-linear processes that depend on higher order moments. Of course this is an expected utility dominance concept. One prospect dominates the other if its expected utility is higher. But which member of a class of utility functions? Or, which distributions? Can dominance hold over entire classes of utility functions (uniformly)? This expected utility definition is not actionable from a statistical point of view, but there are equivalent conditions on CDF and/or quantiles. So, it becomes approachable and testable based on a sample of data. You can imagine it graphically. If one cumulative distribution falls everywhere to the left of the other, it is first order dominant. And for risk (error loss) it's the other way around. McFadden and others had begun to look at it. Perhaps not so ironically, it's Lehmann who has the first rigorous

paper on this, referring to it as comparison of experiments. He's really the first expositor of uniform statistical dominance over classes of risk functions. Later on scholars concerned with risk and returns took it on. And two of my colleagues at Southern Methodist, Hadar and Russell are inventors, simultaneously with H. Levi, of second order dominance.

Different orders of dominance have to do with classes of utility functions with increasing concavity and more restrictive attitudes to risk (or inequality). The large class of increasing and concave preferences covers almost all the preference functions we use in science. Then you have this paper by McFadden that was not ever published in a journal. It was published in a book in honor of Joe Hadar, where he compares these prospects and actually gives a test, the Kolmogorov-Smirnoff supremum test. It's a very conservative test of dominance. His first paper had two significant restrictive conditions. He assumed samples are IID. And that the two prospects you're comparing, x and y, have to be independent. That I would argue nullifies the practical usefulness of dominance for policy evaluation and treatment effect analysis. Prospects are never really independent. Because it's typically in settings that you use these, the two conditions, the two situations you compare are highly related. You have an economic event or policy, and you say what happened to the income distribution? It's the same people's incomes before and after a policy. So the idea that x and y are independent is a non-starter. Thus the 1989 paper in that book is a fantastic starting paper. The idea to use the Kolmogorov- Smirnov test, and use it both ways simultaneously to identify which distribution dominates. So I had heard that he had a working paper in which he does it for financial returns, in which he relaxes some of those assumptions. I couldn't get my hands on this paper, so I actually travelled to Berkeley. Made a date with him, we went to lunch, and he gave me a copy of that paper. It turns out it had been up for revision for Econometrica, but never revised. It had some technically incomplete sections in it. I started using the Kolmogorov-Smirnov test of Dominance many years before I had a rigorous theory for it, I used naïve bootstrap since there was no asymptotic distribution developed for these things. And there is a problem in this setting because these tests are not pivotal. They actually depend on the unknown CDFs. It couldn't be pivotalized. So I was toying with doing bootstrap resampling distribution estimation of these tests in a context that is not pivotal. Going back to my PhD, and Denis Sargan's work on resampling (Sargan 1982, Advances in Econometrics), and how resampling methods estimate higher order approximations, I was confident that bootstrap would work. There are better approximations, however, based on subsampling and others.

Why do you think it works better? Is it pivotal?

When journal editors ask you to prove the validity of bootstrap that's what they mean. Show me convergence to the asymptotic distribution. But if the asymptotic distribution is good enough why bother with bootstrap? That's my position. [LAUGH]. I accept that every time you use an approximation, you have to show something about its performance and limiting "validity", so I accept that. But it is not satisfactory.

So that's a lot of digression from stochastic dominance. I started using bootstrap implementations of these stochastic dominance tests. And Oliver Linton called upon me and showed me some proofs in which he could actually derive the asymptotic distribution of subsampling and bootstrapped versions of these tests, for any order of dominance. Later we asked Yoon-Jae Whang to join us, and he tightened up some of the theorems. That led to the Linton, Maasoumi and Whang (2005) paper in which we also allowed for dependent data like time series, and for prospects that were not independent. We relaxed all of the restrictive assumptions that were in the original McFadden papers.

So why did Linton take an interest?

Linton, remember, had been a student of McFadden at Berkeley. So he must have been aware of the test. So Oliver and I worked on that after a presentation by me at the LSE in which I proposed bootstrapping instead of simulation that Dan McFadden had proposed. Oliver quickly thought subsampling was the way to go. McFadden had simulation-based suggestions for exchangeable processes. My work showed that these approximations work well, but I didn't have the proofs. Later we also added re-centered bootstrap methods and put it in the paper to satisfy Joel Horowitz, Econometrica co-editor (the referees were fine and very supportive!). I think Joel was offended by my critical comments on "Econometric society barons" and the inadequacy of "first order" asymptotic approximations (Fellow's opinion, Journal of Econometrics). After the paper appeared in RES, we realized several pages and tables had been inadvertently omitted by the printers! So a corrigendum was published later with those missing parts. And so, that's the history of that paper with Oliver. Yoon-Jae Whang has a super new book on stochastic dominance tests that I strongly recommend ("Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications", Cambridge University Press).

Are your tests valid for residuals and how are they related to other tests?

This is important from an econometric point of view: Our tests are developed and valid for residuals of models. That's important because one may rightfully question whether income differences, say, are due to education. It could be because of age or experience. So one can control for covariates. And what is left is the thing that may be actionable, or not.

The models we allow would be considered pretty generic. Only their parameter estimators have to admit certain common expansions. Once you have that, our tests are in fact shown to be valid for the residuals. Style analysis in finance is a good example, where people don't want to just compare returns, they want to control for certain factors of the market. In finance, almost all of the extensive uses of SD are informal or graphical. No degrees of confidence are reported, as could be done with our tests. I should say that there are competing tests by Barret and Donald, Gordon Anderson, and a small group of researchers. I wrote a little chapter on the history of these multiple inequality restriction tests for Badi Baltagi's edited book, A Companion to Econometric Theory[26]. The problem of testing multiple inequality restrictions had been tackled by Frank Wolak, and going back to Robertson, and others, based on order statistics. Gordon Fisher and his student Quan Xu had also developed an SD test based on these earlier quantile conditions. Basically, one gets to use what is known as the Chi-Bar squared distribution, even under assumption of normal processes. The tests are difficult to implement and have lower power than the Kolmogorov-Smirnov tests.

There is still no smoothed nonparametric kernel estimation of the component CDFs in these tests, right?

What a beautiful question, and it's kind of your cup of tea as a leader of NP econometrics! I should throw it in your lap as one of those things that are left to do. So in all of these tests one has to estimate CDFs, or cumulative CDFs and integrals thereof. The favored estimate is the empirical CDF. There is no smooth estimator. I believe that has not been done. There are smooth NP estimators for CDFs, such as those discussed in the book by Qi Li and Jeff Racine[27]. Now we know a lot more. Actually we have used such estimators for SD tests, empirically, with Jeff Racine. The book by Whang has some discussion on this, and also of conditional SD tests.

---

[26] Baltagi, B. H., & Baltagi, B. H. (Eds.). (2001). *A companion to theoretical econometrics*. Oxford: Blackwell.

[27] Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Well, which paper is with Jeff?

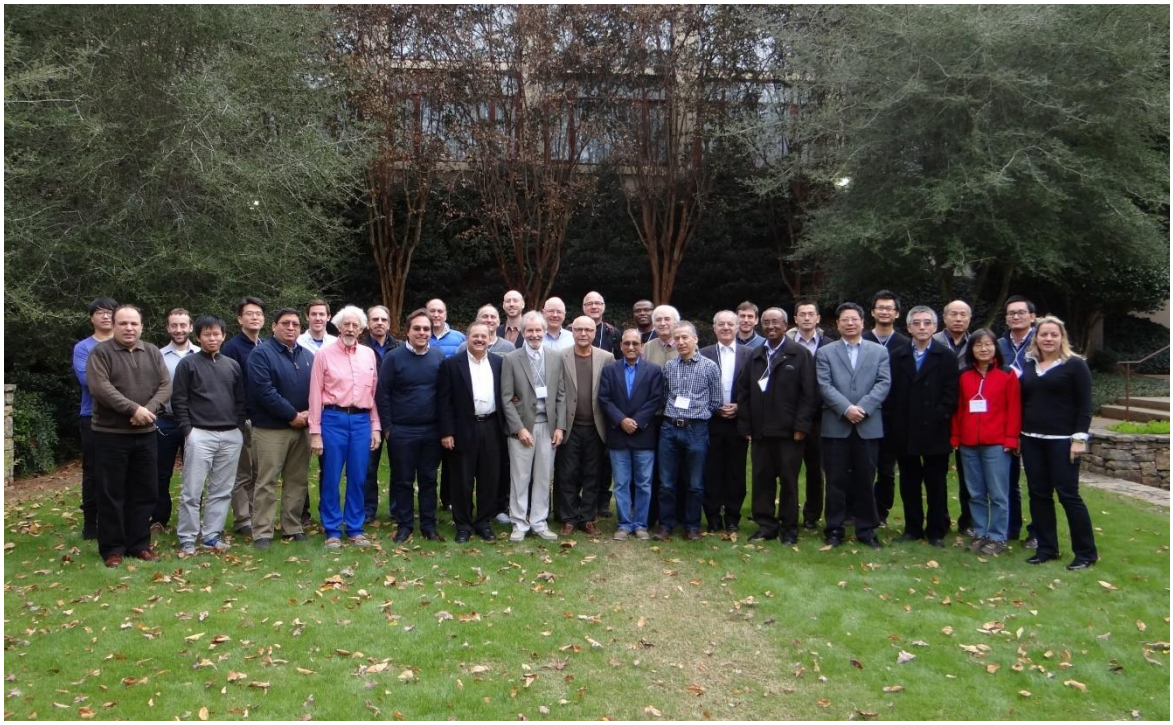
The multivariable poverty paper, Maasoumi and Racine (2016, JoE)[28]. We have joint conditional distributions of income and health, for different education levels and ethnicities. We aggregate these dimensions in an interesting way that audiences have received very well. The joint nonparametric distribution is projected unto quantile (equi-probable) sets first. Then we fit my well-being aggregator functions to these quantile sets to derive the weights to different dimensions and substitution between them. Having obtained these aggregate wellbeing functions, we do dominance testing for different values of conditioning groups and levels. This is one of my favorite papers. [LAUGH] It's related to multidimensionality, the challenge of aggregating.

While we are on this topic, I am curious to know how your work relates to the work in development and labor economics, for example, the extensive work by Jim Heckman who participated in a conference in your honor in 2014 and the work done by Tony Shorrocks.

Tony Shorrocks' work on ideal inequality measures is fundamental. His works are so clear and consequential. My own work on mobility measures is very much a generalization of his ideas about the equality enhancing mobility measures, the so called Maasoumi-Shorrocks-Zandvakili measures. He also has very illuminating and penetrating thoughts and work on uniform ranking of distributions, and the relation between generalized Lorenz and second order stochastic dominance orders.

Jim Heckman is arguably the most rigorous and consequential thinker in modern labor economics. His work on education, human capital, program evaluation, early childhood development, etc. is deep, statistically rigorous, with unusually careful attention to the sampling issues such as sample selection and identification. He is also an early proponent of distribution centric examination of outcomes. My students and I used to feel pretty lonely in our push for a distribution centric evaluation of outcomes until relief came in a big way from Jim, in his call for removing "the veil of ignorance" which was a critique of "average treatment effect" literature in which mean treatment effect of programs and events was the exclusive object of interest. The heterogeneity of outcomes is where the real action is, and now we have some of the best work in

[28] Maasoumi, E., & Racine, J. S. (2016). A solution to aggregation and an application to multidimensional 'well-being' frontiers. Journal of Econometrics, 191(2), 374-383.



Emory University, Conference in Honor of Essie November 15-16, 2014

that area and elsewhere being done on quantile effects, interesting functions of the distribution of outcomes, like Lorenz curves and inequality, entropy functions, etc. We are now capable of identifying the entire distribution of counterfactuals by inverse probability methods, and by inverting estimated conditional quantile functions. In the older days one could only do "decomposition analysis" at the mean only, like Oaxaca-Blinder type decompositions. We are learning, and this was noted by Heckman and few others, and in our recent paper on the gender gap, that quantile effect-based measures of treatment have big challenges in terms of their requirement of "rank invariance" which is not likely satisfied in practice. This means that it is currently better to examine gaps between distributions based on "anonymous" measures of each distribution obtained separately.

Could you please let us know about your joint work on the above with your colleagues and students such as Zandvakili and others?

Earliest joint works on multidimensional wellbeing were with a student at IU, Jin Ho Jeong, and later with a colleague from USC, Jerald Nicklesburg. The work with Jin Ho, Maasoumi and

Jeong (1985, EL)[29], was an attempt to use many indicators available for national wellbeing, like GDP, health, number of radios, paved miles, etc., to produce world inequality measures. I think this is the first of its kind using modern aggregation and generalized entropy inequality measures which I had proposed in the Econometrica 1986 paper. This application actually came out in the Economics Letters, 1985.

The paper with Nickelsburg (1988)[30] appeared in the JBES and was based on Michigan PSID panel data, and reported the intricacies of multidimensional wellbeing and inequality jointly for income, education and health. In that paper we also computed some Nagar like approximations to Theil's entropy inequality measures. These expansions were done with Hans Theil in the late 70s when he was a regular visitor to USC. They show the dependence of entropy functions on higher order moments of distributions, like skewness and kurtosis. Often, I am asked what one should expect from using entropies over variance (!). Well the answer is that entropy is a function of all the moments of a distribution, and these small-sigma expansions expressly reveal the relation to the first four moments. In the above-mentioned paper with Nicklesburg we noted that these expansions were poorer as approximations the larger the skewness and kurtosis. This is an ironic property of moment expansions and finite sample approximations: they do rather poorly when distributions differ widely from normal! This is not a comparison with first order asymptotics which would have us believe everything is approximately Gaussian (to a first order of approximation)!

The work with Zandvakili[31] is on mobility measures which is actually related to the multidimensional well-being work. If one considers an individual's income at different ages or points in their life cycle, say M periods, then the individual's "permanent income", one that determines her consumption decisions, savings and investments, and smoothing, substitutions, etc. over M periods, would be the relevant multidimensional measure of her "income". How does a social evaluation function, say one that underlies an inequality measure of income distribution, evolve as one increases M? If multi-period income inequality declines compared with instantaneous (yearly, say) inequality, we have "equalizing mobility". This is a generalization of

[29] Maasoumi, E., & Jeong, J. H. (1985). The trend and the measurement of world inequality over extended periods of accounting. Economics Letters, 19(3), 295-301.

[30] Maasoumi, E., & Nickelsburg, G. (1988). Multivariate measures of well-being and an analysis of inequality in the Michigan data. Journal of Business & Economic Statistics, 6(3), 327-334.

[31] Maasoumi, E., & Zandvakili, S. (1986). A class of generalized measures of mobility with applications. Economics Letters, 22(1), 97-102.

a central notion of mobility embodied in Tony Shorrock's mobility measures in which he considered arithmetic averages of incomes over time. Zandvakili and I considered finite substitution in very flexible aggregators that include arithmetic average and other linear weighted averages that assume income is infinitely substitutable at any age. In linear aggregation, a dollar of income is as "good" when you are 18 as when you are an accomplished 60-year-old! Our mobility measures were perhaps the first that had welfare (decision) theoretic bases for why mobility is a good thing. The panel data studies of mobility effectively consider high variance as a measure of mobility. It is not clear why variance is a "good thing", as in high rates of layoffs for well paid workers in one period and high wages in the next (as an example). In effect, our mobility measures are comparisons of equilibrium income distributions. Lorenz curves and generalized Lorenz curves are the way to make such comparisons, and that is what we do in effect. We have reported mobility profiles for the US and Germany based on members of these mobility measures. For instance, I think the Maasoumi and Trede (2001)[32] paper is the first to find Germany was more mobile than the US over several decades, a point generally lost in panel data econometric studies.

Have you looked into causality issues in your work, and what do you think about challenges and directions in which work in this area is going?

Entropy notions of dependence allow a definition of statistical causality in terms of the fundamental laws, the joint conditional distributions. This is a major improvement over causality definitions that are variance based. If variance is a measure of uncertainty and lack of information (it is not a good measure most of the time), then does conditioning on some additional variable(s) reduce the variance/uncertainty/predictability? You can see that these notions of Granger/Sims causality are fundamentally limited to a Gaussian world, and higher order uncertainties are not well represented by the variance. The notion of "information gain", defined over two competing conditional distributions, say, goes far beyond the variance. It provides a full entropic measure of information and uncertainty. The copula is a useful way of seeing how two variables are informative about each other, given others, as we discussed earlier. Gouriéroux, Monfort and

[32] Maasoumi, E., & Trede, M. (2001). Comparing income mobility in Germany and the United States using generalized entropy mobility measures. Review of Economics and Statistics, 83(3), 551-559.

Renault[33] had a beautiful paper on this topic of causality based on information theory concepts in the 80s. I think it was finally published in *Annales d' Économie et de Statistique*.

To be fair, there are implementation challenges in many dimensional settings. For instance, Granger causality can be estimated based on conditional variance reductions with a very large set of historical observations, conditional on a long time series of past observations. This is not really possible, straightforwardly, if one insists on nonparametrically estimating joint distributions. There are moment expansion formulae for entropic objects that can be estimated, however. This may follow what I did with Hans Theil for entropy inequality measures, using "small sigma" expansions which were shown to be functions of such higher order moments. Semi nonparametric and "big data" regularization shortcuts have begun to be examined that are promising. A paper was presented on this at the IAAE in Cyprus (2019), I am embarrassed that I don't recall the title or the author!

Causality is a major challenge, conceptually and statistically. There is an ongoing debate, a renewed debate really, about causal models in the treatment effect literature, with the potential outcome paradigm. It is intermingled with inference issues, model misspecification, nature of data (natural experiments, randomized experiments, and observational data, etc.). I actually tried my hand at a taxonomy of these challenges in the early 80s (1984, Economic Education book by William Becker) in which I summarized these challenges to generalizability and internal validity. Here is an area in which other fields have developed more than economics, as evidenced by the fundamental writings, like Rubin's work, that have inspired the potential outcome approach in (mostly) microeconomics and econometrics. I think it is fair to say that this area has done more to highlight and expose the nature of otherwise implicit assumptions that underlie claims of causality and partial effects when people run regressions and use models for policy and forecasting. Efforts at robust model specification and inference notwithstanding, the treatment effect area instructs us well as to how tough it is to claim causal/partial effects. Model misspecification and required assumptions for identification are very challenging. We are learning the value of focusing on other "feasible" objects, like pseudo-true effects, and their limitations! The new "big data" and machine learning methods help in some directions but raise similar issues with respect to identification of causal effects vs mere predictions and projections.

[33] Gourieroux, C., Monfort, A., & Renault, E. (1987). Kullback causality measures. *Annales d'Economie et de Statistique*, 369-410.

Since you have been a distinguished scholar who has dedicated his work on studying distributions and modeling it would be good to know about your opinions on nonparametric kernel distributions and modeling, and recent developments in machine learning procedures like random forests and boosting. Can they play a big role, especially for high dimensional econometrics?

Big data approaches and machine learning are here to stay. We have all been "data scientists" but didn't know it! So is nonparametric inference. But there is a tension here between good prediction vs mechanism discovery. If one has 150 indicators (regressors), the degree of multicollinearity (cofactor effects) is too high to allow identification of ceteris paribus effects of each factor. But we know certain functions are perfectly well identified, such as the conditional mean and similar functions (R-squared and residual functions are unaffected etc.). This is one secret of "big data" methods. We can handle a lot of variables as long as the goal is to estimate the conditional mean and certain other functions of the parameters. Penalization methods like LASSO type approaches don't identify partial effects, really, and incorporate a lot a priori restrictions. Mostly we start with the prior that every partial effect is zero (or negligible) until proven otherwise by the data! The second secret of big data methods is the revolution in the processing speed of computers. It's not that we have found new ways of doing things. There are things we could not do fast and cheaply enough that we can now do. Numerical searches and optimization algorithms have made previously tedious methods feasible and relevant. This is the big deal. Data scientists are (hopefully still good) statisticians who can employ processor intensive numerical algorithms. Mechanism discovery is taking a back seat, and prediction/fit is king. But I am one of those who believe these efforts provide new observations on "the apple falling" and provide us with new insights when we go back to the drawing board to model mechanisms. A financier wishing to get the best return from investments may merely care about the best predicting black box. Policy makers will need deeper and actionable knowledge of the underlying mechanisms.

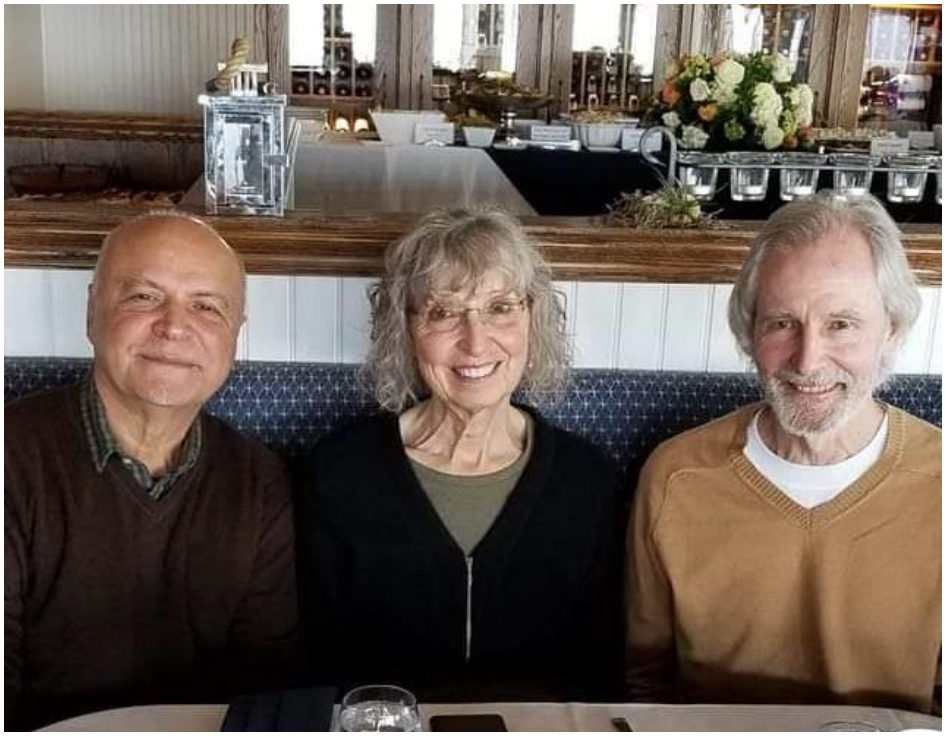
Nonparametric methods are very attractive in low dimension situations. Semi-nonparametric approaches are also promising, when pure nonparametric smoothing methods are not feasible. But it is very difficult to interpret their findings. There are no unique functional and polynomial sets for these methods, and the number of terms is a challenging "tuning" parameter with possibly large consequences in terms of finding partial effects. The book by Qi Li and Jeff Racine on nonparametrics is a modern classic, and so is your book with Pagan. They continue to teach generations of students and scholars. I use these methods as often as is practical.

I should say that in multidimensional representations of "latent objects", like "well-being", "risk", "happiness", health status, etc., big data methods can be very promising. Instead of a priori decisions about indicators of well-being, say, like income, health (and its many indicators), education, we can "aggregate" in statistically meaningful and reasonable ways all of these informative indicators. This was proposed in my 1986 paper on multidimensional measures. I think I mentioned elsewhere that we are now extending this idea of "latent objects" to Data Generating Processes (DGPs). Every proposed model is viewed here as an "indicator", and proxy, and their aggregates represent a better approximation to the latent DGP. This subsumes "meta-analysis", model averaging, and big data when it is mostly variable selection algorithms based on linear or neural nets, and similar filters.

To your question, sample splitting and subsampling designs, including various bootstrap resampling methods, are all used in this context to implement data driven solutions, and to allow training samples for out of sample optimization and even inference. But I do think "inference" is currently very confused, especially with respect to objects or targets of inference. Each model is a poor object, perhaps even a misguided target of inference when all are misspecified approximations. What is coefficient (impacts and derivatives) supposed to converge to? We know they converge to something, commonly referred to as pseudo-true objects. But what are these objects other than statistically well-defined entities. They are also not unique since they can be defined as optimizing values relative to estimation criteria or prediction objective functions.

Within this setup how you would describe your collaborations with Jeff Racine on nonparametric related papers.

Jeff has been a long-time collaborator and coauthor on several projects that involve nonparametric estimation. He has been one of the many ambassadors that you have produced in econometrics, and who have joined with me in research and friendship. There are two broad areas of collaboration with Jeff that reflect the two projects we have talked about: non-linear models and dependence notions based on entropies and determining weights and substitution degrees in multidimensional measures of well-being. Based on the work we needed to do with general entropic measures of dependence, Jeff went on to utilize his highly praised and appreciated computing and software expertise to write a famous package in R, called NP. The techniques



Essie, Deb, Peter at Yale's Conference in Honor of Peter Phillips, October 2018

available in this package allow computation of nonparametric estimation of multivariate distributions, estimation of generalized entropies and, our favorite, the Hellinger measure!

I should add a sentimental note here about Clive Granger: He was present and instrumental in the development of a normalized Hellinger measure as an entropic measure of distance between distributions. He and I developed these ideas when I was visiting UCSD in 1991, and Jeff was also a visitor then. We wrote a paper together, the three of us, on the first implementation of these ideas to tests for nonlinear "serial" dependence.

You have done extensive work in econometric theory and have been associated with Peter Phillips as a colleague, friend, and co-author. We would like to hear your comments.

Surely there won't be enough space here to cover this large topic for me. Peter is undoubtedly the biggest influence on my work in econometrics in ways that go beyond specific contributions. I have covered some of this in our prior discussions. But his early work and extensive interactions between us in the late 70s and early 80s were inspiring and instructive for me. I read everything he was writing and discussed with him. He was really ahead of everyone else on exact distribution

theory. We had a joint NSF grant, rather a generous one for a number of years. Only one of the pieces that we worked on went on to be published, Maasoumi and Phillips (1982) that has received some attention, again, in recent times because of its results and observations on what happens to the distribution of IV estimators under misspecification. We also showed that there was a lot of contemporaneous confusion about "response surface" regressions and limits of what you can discover from Monte Carlo simulation studies. It was quite an interesting collaboration. We often found, and were held back by, mistakes in otherwise well-known published works! It was quite disheartening.

There were rumbles in the community about the impact of and limits to exact finite sample distribution work. This was not unfair as much was limited to linear models and Gaussian processes, and still rather beyond the grasp of many, and not a little bothersome for those who wanted to push on with empirical work in more complex models. It was about this time period, late 1970s, that we began to have an explosion of first order asymptotic approximation and its application to inference in much more complex modelling and data situations. I would say that I am one of the few who still has doubts about the efficacy of asymptotic theory to adjudicate inferences. I think we have become rather numb to the notion that everything is asymptotically Gaussian and its normalized squares is a Chi Squared. I have complained about this, loudly, and in print (*"On the Relevance of First Order Asymptotic Theory to Economics*," *Journal of Econometrics*, 2001)[34]. I have an uneasy feeling that this imprecise method of approximation may be partly to blame for a certain, tangible lack of respect for "inference" and model performance amongst some empirical economists, especially macro researchers who show a clear degree of cynicism and disregard for rigorous statistical inference, and even training of students in these areas. An example of this is the attitude that "my model is my toy model and I am sticking to it". Tom Sargent confirms the evolution of this kind of attitude in his interview with Bill Barnett (Inside the Economist's Mind[35]), when he recounts reactions of leading macro scholars to almost universal rejection of their models by econometric tests!

[34] Maasoumi, E. (2001). On the relevance of first-order asymptotic theory to economics. Journal of Econometrics, 100(1), 83-86.

[35] Samuelson, P. A., & Barnett, W. A. (Eds.). (2009). *Inside the economist's mind: conversations with eminent economists*. John Wiley & Sons.

Peter is surely at the forefront of scholars who have taken the standards of rigor to another level. I did not participate in the "unit root and non-stationarity revolution" for philosophical reasons. But his leadership in that area is legendary and well known.

On another note, Peter has always cared deeply about development of the sciences and training, and especially younger scholars. He and I spent many days and wonderful hours talking about Econometrica, the difficulty of publishing in econometrics, and naturally, the development of ideas that led to Econometric Theory being launched by Peter and becoming the premier theory outlet in our field. I will have more to say about him later as we have his 40th year at Yale celebrated soon, and many more special issues of journals to celebrate his many contributions and stunning mentorship record, including 100 or more Ph.D. students! Peter has been a great mentor and supporter for me, and a source of inspiration and pride as my best friend since the middle 70s.

What are your views on Bayesian approaches?

"Bayesian" has moved toward frequentist ways, and vice versa. These distinctions have served their purpose and perhaps are exaggerated. Fully frequentist works have Bayesian interpretations and methods, and Bayesian methods in practice get very "empirical" and use "objective priors" which I regard as convenient interpretations and representations of "information". There has always been a rather misleading and artificial separation between variables and parameters. Model parameters are functions of the distribution law that generates the variables and observations. They are transformations of objects (may be parameters) of the joint distributions which represent our a priori theories about additional relations between observables (like moment restrictions). Given these models are speculative by nature, the parameters of models are questionable objects on which to have puritanical Bayesian priors, as in full probability laws on parameters. We have odds on their likely values, and we have (often not mentioned) odds on entire model classes and their members being "correct". We learn from the objective data observations through the questions we raise and theories we examine. There is no such thing as purely data/evidence based learning. I ask my students to consider a simple mean model for n observations, without assuming all the observations have the same mean. We have to posit a priori models for these heterogenous means as functions of other variables with fixed effects. There are many good ways of understanding the role of tentative propositions, the role played by models, "Bayesian priors" and their updating. I

learned from Denis Sargan that it was silly to be dogmatic about these issues. Doing things well, precisely, rigorously, and being transparent are far more challenging.

Also, what are your views on macroeconomics?

This is a sour subject for most econometricians. Most are reluctant to speak publicly about it but are quite forceful in their negative views of modern macro. Most if not all economists choose the discipline because they are moved to understand and help improve the aggregate economy and policy making at that grand level. One may be forgiven to get high on the depth of the questions and challenges of macro and coexist with rather stagnant and pedestrian advances in the economic discipline of macro.

My way of looking at issues that impinge on progress in macro is to focus on identification challenges of macro-econometric models. Paul Romer is not quite right when he says people have not spoken against the problems in the dominant approaches to macro. Many of my early papers, including my Thesis chapters in 1977, were critical of then developing models and methods in macroeconomics. He is right to say that it was and is professionally costly to criticize workers in this area. Let me be concise: in macro there is just one sample of aggregate observations. If the data does not "identify" the objects that a model of this data postulates, that is a problematic mathematical property of the model. It is not accessible to statistical/scientific/objective learning and discovery. To pretend to identify objects in these models by "Bayesian" priors is an abuse of the "Bayesian" attribution. Bayesian updating means that posteriors become the later priors, to be combined with new evidence/samples. But there is no updating of priors in modern macroeconomics, or any alternative samples to learn from. Minnesota priors, I note, have not been updated for several decades. Data snooping in this context is a major problem. The multitude of tests, pretests, on the same data sample, inform the next round of models by successive researchers. This problem cannot be accommodated within the Bayesian statistical paradigm. MCMC sampling and integration techniques are brilliant. But they have little or nothing to do with identifying an under-identified object, or with data snooping.

Modern macro has also replaced allegedly "incredible" a priori restrictions of macro theory, with patently incredible and expedient (and mostly false and unverifiable) a priori statistical assumptions on the "shock" terms. Everything is explained by the "shocks" and un-observables. If the model and observed variables fail miserably to explain things, it must be a set of calibrated

correlations between “shocks” and residuals that “explain” things. Technological shocks, supply shocks, policy shocks! Some of the best schools in the world have clearly decided to “sit this one out” until we get a grip and new approaches to macro emerge.

You have extensively contributed to the econometrics profession in many ways. Could you please let us know your work and experiences as an editor of ER which has developed into a first-rate econometrics journal? Specially, about its history, growth, and future directions.

Some of the ideas discussed with Peter which I alluded to earlier addressed the problems of publishing perfectly good works in econometrics in the leading journals. There were really only two obvious places, Econometrica and Journal of Econometrics. The proposition of a journal like ER came to me early on, as a suggestive idea by Dennis Aigner, my colleague at USC. He had been approached by publishers to start a journal focused on “reviews” of major developments in the field. He asked if I were able and willing to do it. I was flattered but, upon reflection and a moment of rare modesty (!), I decided I was too young for it in 1979-1980! Dale Poirier later accepted to be editor and a wonderful editorial board was in operation for about 5 years with several classic papers and exchanges published in the first few issues. In 1987 the publisher had conducted a large survey and had a short list of candidates to be the next editor in chief. It was Peter Schmidt who called me and guided the process that resulted in my accepting to be editor, with major changes to the format and operations of the journal, including a radical switch to “regular papers”, with occasional major surveys and exchanges, and full refereeing of everything. We had two issues a year, copies of the original papers published by Marcel Dekker. Now we have 10 issues/year, published by Taylor and Francis, and pressed to deal with challenges of 85-90% “rejection” rates of often good papers. Almost every major figure in econometrics has served the Board and helped establish ER as a foundational outlet in econometrics. We have recently recognized our board members and frequently published authors in ER with designation as Fellow of ER. You were the first person I called to invite to the Editorial Board in 1987! ER is a collective achievement of the econometrics profession since 1982.

You have contributed so much to teaching and supervising students. What have been your motivations, and who were the best students among the students you have supervised?

I have always been blessed with a love of teaching and instruction. I was lucky to start early, as a 22-23-year-old "lecturer" at the LSE and later at Birmingham (where Peter hired me after my work permit at the LSE had been denied by the Home Office!). Of all the things that are heavenly about teaching, helping people to do the best they can, sharing the journey especially when the outcome is not perhaps as evident as with the most select groups at the Ivies, is the most gratifying part of my professional life. I only regret I did not do more, and I have lamented this, selfishly blaming other things, in my "Reflections" piece ("*Reflections*," *Journal of Economic Surveys*, 2000[36]).

I suppose one of my most successful students in recent years is Le Wang. We just published a paper on many years of research on the gender gap (JPE, 2019)[37]. But I have had many students who taught me, and many younger coworkers without whom I could not have progressed with my ideas and work. I have also some recent Ph.D. students whom I persuaded to return home, such as to China. Example, Ke Wu at Renmin, who is a very impressive scholar. One of my earliest students, Sourushe Zandvakili, did not want to do econometrics! So I advised him to engage in my then current interest in mobility. That work has had a good impact and is now being revisited with the re-emergence of mobility and well-being as dominant topics in economics and policy. I also enjoy frequent visiting appointments, such as at UCSD, Iran and elsewhere, where I have an opportunity to "infect" students and impressionable young scholars with ideas about IT, well-being, misspecification, poverty, and little missteps and misinterpretations in econometrics (according to me!) I have several successful students serving as senior professors in Australia, such as Denzil Fiebig and Joseph Hirschberg.

What is your opinion about the way the econometrics text books have been developed over the years? Do you see any gap in them with the research and practice of econometrics, and the learning outcomes of students?

I think we have had a challenge with suitable textbooks in this field at all levels. At the graduate level things have improved due to immediate availability and online access to public versions of scholarly papers, and appearance of several important "Advances in Econometrics" volumes,

---

[36] Maasoumi, Esfandiar (Essie), 2000. " Reflections," Journal of Economic Surveys, Wiley Blackwell, vol. 14(4), pages 501-509, September.

[37] Maasoumi, E., & Wang, L. (2019). The gender gap between earnings distributions. *Journal of Political Economy*, *127*(5), 2438-2504.

covering state of the art surveys of leading topics by leading authors. There remains a massive problem of inequity between programs in terms of that tangible, but hard to quantify, motivational teaching, deep learning of ideas and connectivity to prior works and fields, which only the best can provide. Maybe more lectures can be prepared and shared on YouTube and otherwise. This is happening, but not in an organized way, yet. It will happen. There are some NBER video graphed lectures that are very well done, especially on treatment effect and potential outcome paradigm. There are several very good ones on "big data" and machine learning methods. Some leading scholars provide free access to their lectures and code; notable examples are Bruce Hansen who provides a free graduate textbook, and Victor Chernozhokov (MIT) who provides slides and extensive code.

There are a lot more texts at the undergraduate level, but many are not satisfactory, or at least not suited to every group of students. Our biggest challenge is diversity of student backgrounds in formal subjects and formal thinking. This is a major ongoing problem, with no sign of improvement. In a class of 30 students in their third or even senior year, one can find several modes of prior knowledge and command of statistics and basic mathematical skills. This is endemic and perhaps universal. The need for selection in such topics goes against the equally compelling considerations of equal opportunity and inclusivity. We are all subject to the same phenomenon: first degrees spend a lot of time covering what is no longer really covered in high schools, especially as far as STEM subjects are concerned. The demand on all of us to just teach them "how to use" techniques, and run the code and produce the tables, is very strong. Personally, I don't see a proportionate rise in the percentage of deeper thinkers and truly knowledgeable experts growing with the massive increase in scale of education everywhere. Online learning and teaching, around the corner with scale, is pertinent to these observations and raises deeply troubling, and yet exciting questions. Imagine the democratic and equitable availability of the best lecture on IV estimation, say, by the best presenter in the world, available to all on their laptop. One can only salivate. I explored this prospect with a few former students in the mid-80s. We were defeated by technological challenges of producing high quality DVDs and video files that were efficient enough and easily accessed!

Bipolar development in econometrics: very sophisticated and advanced techniques for inference and model evaluation, on the one hand, and prominent empirical work that is more

accommodating to undoubtedly massive new pressures for immediate relevance and impact. Mostly harmless econometrics seems mostly harmful for evidence-based progress.

Teaching has also become polarized: best practice and high-level training at best schools and programs, and the lag in adjusting to new knowledge fast enough. Great advances have taken place in robust methods, big data, identification in less restrictive models without additivity and separability, NP methods and shape restrictions. Computing power and methods pose questions about the role and primacy of older mathematical methods which prioritized derivation of closed form solutions. Can our introductory textbooks be written to accommodate this computational revolution? Should they?

How do you feel to be the Fellow of several organizations?

Naturally honored by it all. Some are from automatic qualifying processes of which I am very fond! Others are gratifying and deserved by far more people than the few who receive them. In economics we don't have enough means of recognizing people, so the few ways that we have available have acquired unreasonable weight and status. My reaction to a negative comment about the latest batch of Fellows of Econometric Society, for instance, is that they all deserve it, and many more that didn't get recognized. I am not aware of any Fellow of ES that did not deserve such a small recognition by another 400 or so economists (mostly non-econometricians!). People complain about the dominance of major universities in these matters. I am not surprised by it, certainly not shocked. It wouldn't do to consult Yellow Pages when you are looking for a good doctor! Of course we can do better. Jim Heckman and a coauthor just published a negative assessment of the "top five" journal status and assessments of publications. But this may well be more a "top 10" schools' problem, for which we all share the blame. Other disciplines, while not free of these issues, are less extreme than economics.

You have worked with so many distinguished scholars like Theil, Granger, Phillips, Diewert. Could you please describe your experiences working with them, and who has made a lasting impact on you?

I consider myself as one of the luckiest workers in our profession in this regard. Starting with the greatest advantage of being groomed at the best econometrics school (LSE) in the world (!), and later on, I have always been treated with special care and enormous generosity of the greatest

minds in our profession. Some were teachers and advisors, like Sargan, Sen, Morishima, Gorman, Durbin, Stuart (of Kendall and Stuart fame, the best "teacher" of the lot!), Hendry and others. Some were undergraduate teachers and teaching assistant for our classes, like Lord Desai (applied econometrics), Lord Dasgupta and Sir Steve Nickell. But my good fortune continued strongly, with Peter Phillips (a slightly more advanced fellow student at the LSE), Henri (Hans) Theil, Clive Granger (both acted like a friend and father to me), Arnold Zellner, Jim Heckman, Ted Anderson, Peter Schmidt, Jerry Hausman, …. the list goes on. This fellow called Aman Ullah hasn't been bad to have on my side all these years, since about two days after graduation, meeting in Vienna! They all work differently, but share one thing in common: an uncommon grace and generosity to others and willingness to engage with others and share. They all set a high bar for me. It is the inequity of exposure to such greatness that I complain of when I assess equality of opportunity in education and human capital.

While this interview is going on as you know a special issue of ER was completed in your honor and 25 distinguished scholars around the world have contributed in it. Do you have any comments on this?

What can I say that would adequately express my gratitude and humility in the face of such a deeply moving response by my beloved friends and colleagues? I am specially honored that Peter Phillips and yourself initiated and coedited this special issue. I am not sure I have been moved as much by any other honors that have come my way. I dare say I feel sad about the absence of my departed friends and mentors like Sargan, Durbin, Zellner, Theil, Anderson, Gorman, and others. I miss them a lot. I now have to work hard to justify this remarkable and meaningful gesture by so many stars in our profession!