Mathematisches Forschungsinstitut Oberwolfach

Report No. 19/2013

DOI: 10.4171/OWR/2013/19

Mathematical Statistics of Partially Identified Objects

Organised by
Victor Chernozhukov, Cambridge MA
Wolfgang Härdle, Berlin
Joel Horowitz, Evanston
Ya'acov Ritov, Jerusalem

21 April – 27 April 2013

Abstract. The workshop brought together leading experts in mathematical statistics, theoretical econometrics and bio-mathematics interested in mathematical objects occurring in the analysis of partially identified structures. The mathematical core of these ubiquitous structures has an impact on all three research areas and is expected to lead to the development of new algorithms for solving such problems.

Mathematics Subject Classification (2010): 62G08, 62G05, 68T05, 68Q32, 62P20, 60D05, 60E15.

Introduction by the Organisers

The workshop "Mathematical Statistics of Partially Identified Objects" was organized by Victor Chernozhukov (Cambridge MA), Wolfgang Härdle (Berlin), Joel Horowitz (Evanston) and Ya'acov Ritov (Jerusalem) and was attended by 23 participants. The program included 21 talks of 60 minutes each, including discussions.

The workshop brought together mathematical statisticians, theoretical econometricians, and bio-mathematicians to understand and further develop the mathematical core of partially identified objects or structures. Partial identification is of ubiquitous nature in the analysis of structural models. Such analysis is relevant in many fields of applied mathematics. For example, when data are generated as outcomes of optimal discrete choices, moment inequalities do not identify the parameter, but they can be highly informative about it by restricting it to lie in the so called identification region. This phenomenon raises a variety of important

mathematical, statistical, and computational problems that were explored in the workshop.

Partial identification also arises in clinical trials. If some subjects do not take the treatments to which they are assigned or if there is attrition from the trial then the effects of treatment on the outcome variable are not identified unless one makes untestable assumptions about the attrition process. These identification problems are frequently explored through sensitivity analysis. A more thorough mathematical characterization of the described effects can be achieved by computing entire identification regions. Similar problems arise in survey research due to survey nonresponse. In the analysis of survey data, nonresponse is often dealt with by modeling the nonresponse process (e.g., assuming nonresponse is random conditional on observed covariates), but the models are not testable empirically and can lead to highly misleading conclusions.

Another example of partial identification is the prediction of the effects of a policy decision following, say, a clinical trial. Even if the results of a trial are not complicated by attrition or non-conformance to assigned treatments, they do not provide point predictions of what outcomes will occur in the general population. A clinical trial, at best, gives the average effect of treatment on a randomly selected group of individuals with the relevant disease or medical condition. However, once a drug or device is approved, those who receive are not randomly selected. Rather, they are chosen through a complicated and poorly understood process involving advice from medical professionals and the preferences of the patients. Consequently, the predicted effects of the new drug in the population are not identified.

The mathematical core problem is the characterization followed by computation of the identification regions. It is interesting to know whether computable identification regions can be formulated using modern tools from such seemingly unconnected fields of mathematics as the theory of the optimal mass transportation and the theory of random sets. These approaches have recently emerged as powerful tools in identification analysis, replacing earlier more primitive methods based on elementary algebraic manipulations. As a result, today, sharp and highly non-trivial bounds on parameters of games with multiple equilibria can be formulated using the random set theory and optimal transportation methods, substantially improving upon earlier non-sharp identification regions obtained for these models. There are also interesting relations to von Neumann's method of alternating projections and maximum entropy methods.

Another challenging problem is estimating and performing inference on identification regions using available finite data. For example, the properties of the likelihood and other methods for performing inference on functionals of the parameter under partial identification are not yet well understood. Another example is the statistical theory of these methods and the appropriateness of convergence notions for stochastic programs, such as epi-convergence and related concepts, which have been developed in variational analysis and operations research. A number of interesting questions also arises in relation to the failure of conventional inferential

methods, e.g., bootstrap, due to limit distributions of relevant inferential statistics failing to be continuous with respect to the underlying probability measures. A number of methods have been suggested to remedy this failure; discussions around the theory of such methods with the vision of generating solutions that are both theoretically sound and practically relevant were at the center of the workshop.

Workshop: Mathematical Statistics of Partially Identified Objects

Table of Contents

Abstracts

Specification Tests for Partially Identified Models defined by Moment Inequalities

Federico A. Bugni
(joint work with Ivan A. Canay, Xiaoxia Shi)

This paper studies the problem of specification testing in partially identified models defined by a finite number of moment equalities and inequalities (henceforth, referred to as (in)equalities). The model can be written as follows. For a parameter vector (θ, F), where $\theta \in \Theta$ is a finite dimensional parameter of interest and F denotes the distribution of the observed data, the model states that

$$E_F[m_j(W_i, \theta)] \geq 0 \text{ for } j = 1, \ldots, p \ ,$$
$$E_F[m_j(W_i, \theta)] = 0 \text{ for } j = p+1, \ldots, k \ ,$$

where $\{W_i\}_{i=1}^n$ is an i.i.d. sequence of random variables with distribution F and $m : \mathbb{R}^d \times \Theta \to \mathbb{R}^k$ is a known measurable function. This model is *partially identified* because the sampling process and the maintained assumptions restrict the value of θ to a set, called the *identified set* and denoted by $\Theta_I(F)$, which is smaller than Θ but potentially larger than a single point.

The model is said to be *correctly specified* (or statistically adequate) when the moment (in)equalities hold for at least one parameter value, i.e., when the identified set $\Theta_I(F)$ is non-empty. A specification test takes correct specification of the model as the null hypothesis and rejects if the data seem to be inconsistent with it, i.e.,

$$H_0 \ : \ \Theta_I(F) \neq \emptyset \text{ (i.e. model is correctly specified)} \ ,$$
$$H_1 \ : \ \Theta_I(F) = \emptyset \text{ (i.e. model is incorrectly specified)} \ .$$

This problem of specification testing in partially identified moment (in)equality models has not been directly addressed in the literature, although several papers (e.g. [4], [1], and [2]) have suggested a valid hypothesis test as a by-product of the construction of confidence sets for the parameter of interest θ. This procedure, which we refer to as Test BP (for by-product), rejects the specification if and only if the confidence set for θ is empty, i.e.,

$$\phi_n^{BP} = 1\{CS_n(1-\alpha) = \emptyset\} \ ,$$

where ϕ_n^{BP} indicates the rejection rule for Test BP and $CS_n(1-\alpha)$ is any valid confidence set of θ with an asymptotic confidence size of $(1-\alpha)$.

In this paper, we propose two new specification tests, referred to as Test RC and Test RS, which reject the specification of the model when the test statistic exceeds a critical value. In fact, both of our tests use the following "infimum" test statistic:

$$T_n \equiv \inf_{\theta \in \Theta} Q_n(\theta) \ , \tag{1}$$

where $Q_n(\theta)$ is the criterion function typically used to construct confidence sets for θ, much in the spirit of the popular J-test in (point-identified) GMM models. The difference between our two specification tests lies in the critical value used to implement the test. That is, we propose

$$\phi_n^j = 1\{T_n > \hat{c}_n^j(1-\alpha)\} \quad \text{for } j \in \{RS, RC\} ,$$

where ϕ_n^j indicates the rejection rule for Test j, T_n is the "infimum" test statistic defined in Eq. (1), and $\hat{c}_n^j(1-\alpha)$ is the critical value corresponding to Test j.

We prove that our tests achieve uniform size control (just like Test BP) and dominate Test BP in terms of power, both in finite samples and in the asymptotic limit. These findings allow us to conclude that Test RC and RS dominate Test BP as specification tests for partially identified moment (in)equality models.

We now describe these findings in more detail, starting with results on uniform size control. Under certain conditions, Theorems 4.1, 5.1, and C.3 show that

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{P}_0} E_F[\phi_n^j] \leq \alpha \quad \text{for } j \in \{BP, RS, RC\} , \tag{2}$$

where α is the significance level of the test and $\mathcal{P}_0$ is a relevant set of distributions of the data in which the moment (in)equality model is correctly specified. According to Eq. (2), the three specification tests considered in our paper control asymptotic size *uniformly* in $\mathcal{P}_0$. We now continue with results regarding statistical power. Under certain conditions, Theorem 6.1 proves that for any sample size n and any data distribution F,

$$\phi_n^{RS} \geq \phi_n^{RC} \geq \phi_n^{BP} . \tag{3}$$

In other words, Test RS rejects more or equal than Test RC which, in turn, rejects more or equal than Test BP. Eq. (3) is a finite sample result that immediately implies a weak power ranking among the three tests. Under certain local alternatives, we show that the weak power ranking can become a strict power ranking, even asymptotically. In particular, Theorem 6.2 indicates that under certain sequences of local alternative hypotheses $\{F_n\}_{n\geq 1}$, we have that

$$\liminf_{n\to\infty} (E_{F_n}[\phi_n^{RC}] - E_{F_n}[\phi_n^{BP}]) > 0 . \tag{4}$$

By combining Eqs.(3) and (4), we conclude that the rejection rate of Test RC is higher or equal than that of Test BP for all sequences of local alternatives, and it can be strictly higher for at least some sequences of local alternatives. It is possible to establish a similar type of comparison between Test RS and Test RC.

Given that our tests dominate Test BP in terms of power, we find it important to compare these inferential procedures in terms of their cost of computation. The implementation of Test RC requires little additional work beyond the computation involved in the confidence set construction, just like in Test BP. In this sense, Test RC attains better power than Test BP at almost no additional cost. Thus, we always recommend that it is implemented. On the other hand, Test RS has even better power than Tests BP and RC, but its computation requires a separate resampling procedure. For this reason, we recommend its use when one has serious interest in the statistical adequacy of the model.

From a methodological point of view, our paper derives the limiting distribution of the "infimum" test statistic under drifting sequences of data distributions and provides two methods to approximate its quantiles. These methodological contributions are relevant in problems that go well beyond specification testing. In particular, [3] show that hypothesis tests based on the "infimum" test statistic can be adapted to address a large class of interesting new problems which include, for example, sub-vector inference (i.e. inference on a particular coordinate of a multivariate parameter θ).

References

[1] D.W.K. Andrews and P. Guggenberger, *Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities*, Econometric Theory **25 (3)** (2009), 669–709.

[2] D.W.K. Andrews and G. Soares, *Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection*, Econometrica **78 (1)** (2010), 119–157.

[3] F.A. Bugni and I.A. Canay and X. Shi, *Inference for Functions of Partially Identified Parameters in Moment Inequality Models*, Mimeo: Duke University, Northwestern University, and University of Wisconsin-Madison (2013).

[4] J.P. Romano and A.M. Shaikh, *Inference for Identifiable Parameters in Partially Identified Econometric Models*, Journal of Statistical Planning and Inference **138** (2008), 2786–2807.

Restricted structural equation models and improved bounds for high-dimensional causal inference

Peter Bühlmann

Our goal is estimation of causal effects based on observational data, see for example [6] or [9]. In general, or for the case of a multivariate Gaussian distribution, the problem is ill-posed due to non-identifiability of "causal directions" from the observational distribution. The usual route to address such non-identifiability issues is to derive identifiable bounds of causal effects.

One approach to pursue the latter is as follows. We assume that the data $X_1, \ldots, X_n$ are i.i.d. from an observational distribution P_{obs}. The distribution of P_{obs} is determined by a structural equation model, or equivalently, from a graphical model with a directed acyclic graph (DAG) D and P_{obs} satisfying a Markov property with respect to the DAG D. (For simplicity, we only consider DAGs; they do not allow for e.g. directed cycles). In general, D is not identifiable from P_{obs} but the so-called Markov equivalence class is identified from P_{obs}, assuming the faithfulness assumption, cf. [9]. From the Markov equivalence class, we can derive lower and upper bounds for causal effects, as propagated in [5]. Such an approach is consistent, even in high-dimensional settings where the number of variables p (the number of vertices in the DAG D) can greatly exceed sample size n, i.e. $p \gg n$, but the underlying structure is sparse, see [3], [5], [1] and [12]. Furthermore, it has been successfully applied in biological systems for predicting unseen gene intervention effects in the organisms *Saccharomyces Cerevisiae* (see

[4]) and *Arabidopsis Thaliana* (see [10]). Limitations of such an approach when using conditional independence testing are shown in [11].

One can avoid the identifiability problem mentioned above by assuming additional restrictions for the structural equation model. Three such restrictions are as follows: (i) non-Gaussian error terms [8], nonlinear functions and additive noise [2], and linear Gaussian systems with same error variances [7]. We argue that under such additional assumptions (additional restrictions), causal inference from observational data is much more accurate and powerful than in the general nonparametic or unrestricted multivariate Gaussian case.

References

[1] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[2] P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters and B. Schölkopf, *Nonlinear causal discovery with additive noise models*, In Advances in Neural Information Processing Systems 21, 22nd Annual Conference on Neural Information Processing Systems (2008), 689–696.

[3] M. Kalisch and P. Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*, Journal of Machine Learning Research **8** (2007), 613–636.

[4] M.H. Maathuis, D. Colombo, M. Kalisch and P. Bühlmann, *Predicting causal effects in large-scale systems from observational data*, Nature Methods **7** (2010), 247–248.

[5] M.H. Maathuis, M. Kalisch and P. Bühlmann, *Estimating high-dimensional intervention effects from observational data*, Annals of Statistics **37** (2009), 3133–3164.

[6] J. Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2000.

[7] J. Peters and P. Bühlmann, *Identifiability of Gaussian structural equation models with same error variances*, Preprint (2012) arXiv:1205.2536.

[8] S. Shimizu, P.O. Hoyer, A. Hyvärinen and A.J. Kerminen, *A linear non-Gaussian acyclic model for causal discovery*, Journal of Machine Learning Research **7** (2006), 2003–2030.

[9] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, The MIT Press, 2nd edition, 2000.

[10] D.J. Stekhoven, D.J., I. Moraes, G. Sveinbjörnsson, L. Hennig, M.H. Maathuis and P. Bühlmann *Causal stability ranking*, Bioinformatics **28** (2012), 2819–2823.

[11] C. Uhler, G. Raskutti, P. Bühlmann and B. Yu, B, *Geometry of faithfulness assumption in causal inference*, Annals of Statistics **41** (2013), 436–463.

[12] S. van de Geer and P. Bühlmann, *ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs*, to appear in the Annals of Statistics (2013).

On the Testability of Identification in Some Nonparametric Models with Endogeneity

Ivan A. Canay

(joint work with Andrés Santos and Azeem M. Shaikh)

Instrumental variables (IV) methods have a prominent role in econometrics due to their ability to uncover causal effects in observational studies. Though traditionally parametric in nature, an important literature has extended IV methods to a variety of nonparametric settings. Among these extensions, of particular prominence is the additively separable specification in which for an outcome of interest Y, a

regressor W, and an instrument Z it is assumed that

$$Y = \theta(W) + \epsilon \ , \tag{1}$$

with ϵ mean independent of Z. Under the maintained assumption that the model is correct, [1] showed identification of θ to be equivalent to the joint distribution of W and Z satisfying a completeness condition.

Despite the widespread use of completeness conditions in econometrics, little evidence has been provided about their reasonableness in applications of interest to economists. We note, however, that since completeness conditions impose restrictions on the distribution of the observed data, it is potentially possible to provide such evidence by testing the validity of these assumptions. This paper explores precisely this possibility. Specifically, we study whether it is possible to test the null hypothesis that a completeness condition does not hold against the alternative that it does hold. Such a hypothesis testing problem is consistent with a setting in which a researcher wishes to assert the model is identified and hopes to find evidence in favor of this claim in the data. This setup is also analogous to tests of rank conditions in linear models with endogeneity, where the null hypothesis is that of rank-deficiency.

In this paper we show that, under commonly imposed restrictions on the distribution of the data, the null hypothesis that the completeness condition does not hold is in fact untestable. Formally, we establish that *any* test will have power no greater than size against *any* alternative. It is therefore not possible to provide empirical evidence in favor of the completeness condition by means of such a test. This conclusion is in contrast to the testability of a failure of the rank condition in linear specifications of θ, for which nontrivial tests do exist under reasonable assumptions. Thus, while completeness conditions provide an intuitive generalization of the rank condition in a linear specification, the empirical implications of these assumptions are substantially different in this sense.

We additionally derive analogous results in two other prominent nonparametric models with endogeneity. The first such model follows the specification in (1) with a pre-specified conditional quantile of ϵ assumed independent of Z. The second such model follows a specification in which θ is allowed to depend nonseparably on both W and ϵ, with the dependence on ϵ being monotonic, and all conditional quantiles of ϵ assumed independent of Z. Due to the nonlinear nature of such models, simple, global rank conditions such as completeness conditions are unavailable. For this reason, we instead directly consider the testability of the null hypothesis that identification fails against the alternative hypothesis that it holds. Analogous to our results concerning the testability of completeness conditions, we obtain conditions under which no nontrivial tests exist for these hypothesis testing problems either.

Let $\{V_i\}_{i=1}^n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ and denote by P^n the n-fold product $\bigotimes_{i=1}^n P$. The hypothesis testing problems we study may then be expressed as

$$H_0 : P \in \mathbf{P}_0 \text{ versus } H_1 : P \in \mathbf{P}_1 \ , \tag{2}$$

where $\mathbf{P}_0$ is the subset of $\mathbf{P}$ for which the null hypothesis holds and $\mathbf{P}_1 = \mathbf{P} \setminus \mathbf{P}_0$ is the subset of $\mathbf{P}$ for which the alternative hypothesis holds. For a sequence of tests $\{\phi_n\}_{n=1}^{\infty}$, the corresponding size at sample size n is given by

$$\sup_{P \in \mathbf{P}_0} E_{P^n}[\phi_n] \ .$$

We show that under commonly imposed restrictions on the set of distributions $\mathbf{P}$, the three hypothesis testing problems we examine share the property that

$$\sup_{P \in \mathbf{P}_1} E_{P^n}[\phi_n] \leq \sup_{P \in \mathbf{P}_0} E_{P^n}[\phi_n] \tag{3}$$

for any sequence of (possibly randomized) tests $\{\phi_n\}_{n=1}^{\infty}$ and any sample size n. Equivalently, result (3) establishes that for *all* tests, the power against *any* alternative $P \in \mathbf{P}_1$ is always bounded above by the size of the test. It also follows from such an assertion, that any sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ that controls asymptotic size at level $\alpha \in (0,1)$ will have asymptotic power no larger than α against any alternative. Formally, (3) immediately yields that

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}_0} E_{P^n}[\phi_n] \leq \alpha \implies \limsup_{n \to \infty} \sup_{P \in \mathbf{P}_1} E_{P^n}[\phi_n] \leq \alpha \ . \tag{4}$$

In this abstract we establish the nonexistence of nontrivial tests for completeness conditions. Toward this end, we first need to introduce additional notation and formally define L^q-completeness. Let $V_i = (X_i, Z_i) \in \mathbf{R}^{d_x} \times \mathbf{R}^{d_z}$ be random variables distributed according to $P \in \mathbf{P}$. For $Z_i = (Z_i^{(1)}, Z_i^{(2)})$, with the subvector $Z_i^{(1)}$ possibly empty, let $W_i = (X_i, Z_i^{(1)}) \in \mathbf{R}^{d_w}$, and for $\Theta(P)$ a set of measurable functions from $\mathbf{R}^{d_w}$ to $\mathbf{R}$, consider the condition on P given by

$$E_P[\theta(W_i)|Z_i] = 0 \ P\text{−a.s. for } \theta \in \Theta(P) \Longrightarrow \theta(W_i) = 0 \ P\text{−a.s.} \tag{5}$$

For $1 \leq q \leq \infty$, the distribution P is said to be L^q-complete with respect to W given Z if condition (5) holds with $\Theta(P) = L^q(P_W)$. Here, P_W denotes the marginal distribution of W under P and $L^q(P_W)$ denotes the set of measurable functions from $\mathbf{R}^{d_w}$ to $\mathbf{R}$ with finite $\|\cdot\|_{L^q(P_W)}$ semi-norm. For the special cases in which $q = 1$ or $q = \infty$, P is sometimes simply said to be complete with respect to W given Z or bounded complete with respect to W given Z, respectively. See [3] for further discussion on these conditions.

The definition of the set of possible distributions $\mathbf{P}$ plays a fundamental role in setting up the hypothesis testing problem. We restrict attention to sets of measures $\mathbf{P}$ that have a common dominating measure. Specifically, letting $\mathbf{M}_{x,z}$ denote the set of all Borel probability measures on $\mathbf{R}^{d_x} \times \mathbf{R}^{d_z}$, and defining

$$\mathbf{M}_{x,z}(\nu) \equiv \{P \in \mathbf{M}_{x,z} : P \ll \nu\} \ , \tag{6}$$

for some Borel measure ν on $\mathbf{R}^{d_x} \times \mathbf{R}^{d_z}$, we let $\mathbf{P} = \mathbf{M}_{x,z}(\nu)$. In this setting, we examine the testability of the null hypothesis that the completeness condition fails, and hence, for a given choice of $\Theta(P)$, we let

$$\mathbf{P}_1 = \mathbf{P} \setminus \mathbf{P}_0 = \{P \in \mathbf{P} : (5) \text{ holds under } P\} \ . \tag{7}$$

Defining the null hypothesis in this manner is analogous to testing the null hypothesis of a failure of the rank condition in a linear specification

Theorem 1. *Suppose that (i) ν is a positive $\sigma-$finite Borel measure on $\mathbf{R}^{d_x} \times \mathbf{R}^{d_z}$, (ii) $\nu = \nu_x \times \nu_z$, where ν_x and ν_z are Borel measures on $\mathbf{R}^{d_x}$ and $\mathbf{R}^{d_z}$, respectively, and (iii) ν_x is atomless on $\mathbf{R}^{d_x}$. If $\mathbf{P} = \mathbf{M}_{x,z}(\nu)$, for $\mathbf{M}_{x,z}(\nu)$ as in (6), and $\mathbf{P}_0$ and $\mathbf{P}_1$ are as in (7) with $\Theta(P) = L^\infty(P_W)$, then, for any sequence of tests $\{\phi_n\}_{n=1}^\infty$*

$$\sup_{P \in \mathbf{P}_1} E_{P^n}[\phi_n] \leq \sup_{P \in \mathbf{P}_0} E_{P^n}[\phi_n] \quad \text{for all } n \geq 1 \ . \tag{8}$$

Underlying our arguments is a powerful result originally found in [4], which we restate due to its importance in our derivations. In the statement of the lemma, $\|P - Q\|_{TV}$ denotes the Total Variation distance between probability measures P and Q.

Lemma 1. *Let $\mathbf{M}$ denote the space of Borel probability measures on a separable metric space $\mathbf{V}$. Suppose $\mathbf{P} \subseteq \mathbf{M}$ and that $\mathbf{P} = \mathbf{P}_0 \cup \mathbf{P}_1$. If for each $P \in \mathbf{P}_1$ there exists a sequence $\{P_k\}_{k=1}^\infty$ in $\mathbf{P}_0$ with $\|P - P_k\|_{TV} \to 0$ as $k \to \infty$, then every sequence of test functions $\{\phi_n\}_{n=1}^\infty$ satisfies*

$$\sup_{P \in \mathbf{P}_1} E_{P^n}[\phi_n] \leq \sup_{P \in \mathbf{P}_0} E_{P^n}[\phi_n] \quad \text{for all } n \geq 1 \ . \tag{9}$$

Two constructive points follow from our results, which we believe have potentially important implications for future research. First, having established the impossibility of producing empirical evidence in favor of completeness or identification conditions, our results emphasize the value of alternative arguments for their plausibility. Recent work that addresses this problem includes [3] and [2], who argue in favor of completeness conditions on the basis of genericity arguments. Second, our analysis highlights the significance of both developing inferential methods that are robust to partial identification and of comparing their performance to nonrobust inferential approaches. For instance, it would be important to understand which inferential method is preferable when identification is believed to hold. We hope the results and arguments in this paper provide motivation for addressing these challenges in future research.

References

[1] Newey, W. K. and Powell, J. L., *Instrumental variable estimation of nonparametric models*, Econometrica **71** (2003), 1565–1578.

[2] Chen, X., V. Chernozhukov, S. Lee and W. Newey, *Local Identification of Nonparametric and Semiparametric Models*, working paper (2013).

[3] Andrews, D., *Examples of L^2-Complete and Boundedly-Complete Distributions*, working paper (2013).

[4] Romano, J., *On nonparametric testing, the uniform behavior of the t-test, and related problems*, Scandinavian Journal of Statistics, **31** (2004), 567–584.

Generalized Instrumental Variable Models

Andrew Chesher and Adam M. Rosen

We extend the application of instrumental variable (IV) methods to a wide class of problems in which multiple values of unobservable variables can be associated with particular combinations of observed endogenous and exogenous variables. Like traditional IV models, the class of generalized instrumental variable (GIV) models studied restricts the impact of exogenous variables on a structural relationship and limits the degree of dependence of observed and unobserved exogenous variables. However, in contrast to traditional IV models, in GIV models the mapping from unobservables to endogenous variables need not admit a unique inverse.

This class of GIV models allows for unobservables to be multivariate and to enter nonseparably into the determination of endogenous variables, thereby removing strong practical limitations on the role of unobserved heterogeneity. These models are typically partially identifying although they include as special cases well-known point identifying IV models. We draw on results in random set theory, e.g. [1], [5], and [4], to provide a characterization of the sharp identified sets delivered by GIV models using distributions of certain random sets in the space of unobserved heterogeneity.

Important examples include models with discrete or mixed continuous/discrete outcomes and continuous unobservables, and models with excess heterogeneity where many combinations of different values of multiple unobserved variables, such as random coefficients, can deliver the same realizations of endogenous variables. In previous papers such as [2] and [3] we have given some results for particular cases in all of which outcomes are discrete. Here we present a complete development and results for a general class which includes problems in which outcomes may be continuous or discrete. We demonstrate the application of our analysis to a continuous outcome random coefficients model with endogeneity and a model with endogenous censoring of a continuous variable.

References

[1] Z. Artstein, *Distributions of Random Sets and Random Selections*, Israel Journal of Mathematics **46**(4) (1983), 313–324.

[2] A. Chesher, A. M. Rosen, K. Smolinski, *An Instrumental Variable Model of Multiple Discrete Choice*, Quantitative Economics **4** (2013), forthcoming.

[3] A. Chesher, A. M. Rosen, *What Do Instrumental Variable Models Deliver with Discrete Dependent Variables*, American Economic Review: Papers & Proceedings **103**(3) (2013), 557–562.

[4] I. Molchanov, *Theory of Random Sets*, Springer Verlag, London (2005).

[5] T. Norberg, *On the Existence of Ordered Couplings of Random Sets - with Applications*, Israel Journal of Mathematics **77** (1992), 241–264.

Significance testing in quantile regression

Holger Dette
(joint work with Stanislav Volgushev, Melanie Birke, Natalie Neumeyer)

Let Y, X and Z denote one-, d and q dimensional random variables, respectively, where Y corresponds to the response and X and Z are the covariates. We assume that the random variables $\{(Y_i, X_i, Z_i)\}_{i=1,\dots,n}$ are independent identically distributed with the same distribution as (Y, X, Z). Let $\tau \in (0,1)$ be fixed. Our aim is to test whether the predictor Z has influence on the conditional τ-quantile of Y, given (X, Z), or whether the variable Z can be omitted. Thus for fixed $\tau \in (0,1)$ we formulate the null hypothesis as

$$H_0 : E[I\{Y \le q_\tau(X)\} - \tau \mid X, Z] = P(Y \le q_\tau(X) \mid X, Z) - \tau = 0 \ a.s., \tag{1}$$

where $q_\tau(X)$ is defined as the conditional τ-quantile of Y, given X, that is

$$P(Y \le q_\tau(X) \mid X) = \tau. \tag{2}$$

It is easy to see that the null hypothesis (1) is equivalent to

$$T(x, z) \equiv 0$$

for all (x, z) in the support of the random variable (X, Z), where the functional T is defined by

$$\begin{aligned} T(x,z) &= E[(P(Y \le q_\tau(X)|X,Z)) - \tau) I\{X \le x\} I\{Z \le z\}] \\ &= E[(I\{Y \le q_\tau(X)\} - \tau) I\{X \le x\} I\{Z \le z\}]. \end{aligned} \tag{3}$$

This functional can be be estimated by the stochastic process

$$T_n(x,z) = \frac{1}{n} \sum_{i=1}^{n} \big(I\{Y_i \le \hat{q}_\tau(X_i)\} - \tau\big) I\{X_i \le x\} I\{Z_i \le z\},, \tag{4}$$

where $(x, z) \in R_X \times R_Z$, R_X and R_Z denote the support of the distributions of the random variables X and Z, respectively, and $\hat{q}_\tau$ is an appropriate estimate of the conditional quantile of Y given X, which will be specified below. A test for the hypothesis of significance of the variable Z for the τ's quantile curve of Y can now easily be obtained by considering a Kolmogorov-Smirnov or Cramer von Mises type statistic based on T_n and rejecting the null hypothesis for large values of this statistic.

Throughout this paper we assume that the sets R_X and R_Z are compact. We will use an approach proposed by [1] who constructed non-crossing estimates of quantile curves using a simultaneous inversion and isotonization of a preliminary estimator of the conditional distribution function $F_{Y|X}$ of Y given X. For this estimator, say $\hat{F}_{Y|X}(y|x;p)$, we will use a smoothed local polynomial estimator of order p, see e.g. [2]. Following [1] we consider a strictly increasing distribution function $G : \mathbb{R} \to (0,1)$, a nonnegative kernel κ with bandwidth b_n, and define the functional

$$H_{G,\kappa,\tau,b_n}(F) := \frac{1}{b_n} \int_0^1 \int_{-\infty}^{\tau} \kappa\Big(\frac{F(G^{-1}(u)) - v}{b_n}\Big) dv du. \tag{5}$$

Because $\hat{F}_{Y|X}$ is consistent, it is intuitively clear that $H_{G,\kappa,\tau,b_n}(\hat{F}_{Y|X}(\cdot|x))$ is a consistent estimate of $H_{G,\kappa,\tau,b_n}(F_{Y|X}(\cdot|x))$. If $b_n \to 0$, this quantity can be approximated as follows

$$\begin{aligned} H_{G,\kappa,\tau,b_n}(F_{Y|X}(\cdot|x)) &\approx \int_{\mathbb{R}} I\{F_{Y|X}(y|x) \le \tau\} dG(y) \\ &= \int_0^1 I\{F_{Y|X}(G^{-1}(v)|x) \le \tau\} dv \;=\; G \circ F^{-1}_{Y|X}(\tau|x), \end{aligned}$$

and as a consequence an estimate of the conditional quantile function $q_\tau(x) = F^{-1}_{Y|X}(\tau|x)$ can be defined by

$$\hat{q}_\tau(x) := G^{-1}(H_{G,\kappa,\tau,b_n}(F_{Y|X}(\cdot|x))). \tag{6}$$

We will consider a generalization of the test statistic T_n defined in (4), where the indicator functions $I\{X_i \le x\}$ are replaced by indicators of more general sets Θ. To be precise let Ξ denote a collection of subsets of $\mathbb{R}^d$ and define $\mathcal{D}_n := \{x \in R_X | [x - h_n\mathbf{1}, x + h_n\mathbf{1}] \subset R_X\}$ (here $\mathbf{1}$ denotes the d-dimensional vector with all entries equal to 1), then all theoretical developments will be based on the statistic

$$T_n(\Theta, z) = \frac{1}{n}\sum_{i=1}^n (I\{Y_i \le \hat{q}_\tau(X_i)\} - \tau) I\{X_i \in \Theta \cap \mathcal{D}_n\} I\{Z_i \le z\}, \quad \Theta \in \Xi, z \in R_Z. \tag{7}$$

The intersection of the sets $\Theta \in \Xi$ with the set $\mathcal{D}_n$ is needed in the theoretical developments to exclude "residuals" $I\{Y_i \le \hat{q}_\tau(X_i)\} - \tau$ corresponding to predictors close to the boundary of R_X. Note that if $\cup_{\Theta\in\Xi}\Theta$ has a positive distance to the boundary of R_X, the collection of sets Ξ_n will equal Ξ whenever h_n is sufficiently small. The following result is proved in [3]

Theorem 0.1. *If the conditions*

- (A1) *The conditional distribution function $F_{Y|X}(y|x)$ is $p+1$ times continuously differentiable with respect to x, y and all partial derivatives are uniformly bounded on $\mathbb{R} \times R_X$. The joint density of (X, Y) is uniformly bounded on $R_X \times \mathbb{R}$. Moreover, $p \ge \max(s, d+1)$.*
- (A2) *The density f_X of the predictor X is $d+1+n_f$ times continuously differentiable with uniformly bounded partial derivatives on R_X and $n_f > d/2$. Moreover $\inf_{x\in R_X} f_X(x) > 0$.*
- (A3) *There exist constants $a, C_1 > 0$ such that*

$$\inf_{(x,y):x\in R_X, |y-q_\tau(x)|\le a} f_{Y|X}(y|x) \ge C_1$$

  where $f_{Y|X}$ denote the conditional density of Y given X.
- (A4) *The function $(z, x) \mapsto F_{Z|X,\varepsilon}(z|x, 0)$ is Hölder-continuous of order $\gamma > 0$ with respect to z and x uniformly in $x \in \mathcal{D}$, i.e.*

$$|F_{Z|X,\varepsilon}(s|x, 0) - F_{Z|X,\varepsilon}(t|\xi, 0)| \le C\|(s, x) - (t, \xi)\|^\gamma_\infty$$

for some finite constant C, where $F_{Z|X,\varepsilon}(z|x,e)$ is the conditional distribution function of Z given $(X,\varepsilon)=(x,e)$.

(A5) $\sup_{x\in\mathcal{D},y\in\mathbb{R},z\in\mathcal{Z}}|f'_{\varepsilon|X,Z}(y\mid x,z)|<\infty$.

(A6) *The class of functions $\mathcal{F}_1=\{u\mapsto I\{u\in\Theta\}|\Theta\in\Xi\}$ satisfies*

$$N_{[]}(\mathcal{F}_1,\varepsilon,L^2(P_X))\le C\varepsilon^{-a}$$

for any sufficiently small $\varepsilon>0$ and a constant C, where $N_{[]}$ denotes the bracketing number [see [4]]

(A7) $\sup_{\Theta\in\Xi}P(X_i\in\Theta,\exists j:[X_i(j)-h_n,X_i(j)+h_n]\not\subset\Theta)=o(1)$ *for $h_n\to 0$.*

are satisfied, then

$$\begin{aligned}T_n(\Theta,z) &= \frac{1}{n}\sum_{i=1}^{n}(I\{\varepsilon_i\le 0\}-\tau)I\{X_i\in\Theta_n\}(I\{Z_i\le z\}-F_{Z|X,\varepsilon}(z|X_i,0))\\ &\quad +o_P(n^{-1/2})\end{aligned}$$

uniformly with respect to $z\in R_Z,\Theta\in\Xi$, where $\varepsilon_i=Y_i-q_\tau(X_i)$, $i=1,\dots,n$,

Corollary 0.2. *If the assumptions of Theorem 0.1 and the null hypothesis H_0 in (1) are satisfied, the process $\sqrt{n}T_n$ converges weakly in $\ell^\infty(\Xi\times R_Z)$ to a centered Gaussian process $\mathbb{T}$ with covariance kernel*

$$\begin{aligned}(8)\qquad k(\Theta_1,y,\Theta_2,z)&=\mathrm{Cov}(\mathbb{T}(\Theta_1,y),\mathbb{T}(\Theta_2,z))=\tau(1-\tau)\mathbb{E}\Big[I\{X\in\Theta_1\cap\Theta_2\}\\ &\quad\times\mathbb{E}\Big[\Big(I\{Z\le y\}-F_{Z|X,\varepsilon}(y|X,0)\Big)\Big(I\{Z\le z\}-F_{Z|X,\varepsilon}(z|X,0)\Big)\Big|X,\varepsilon\Big]\Big].\end{aligned}$$

As a consequence of this result we obtain the weak convergence of functionals such as the Kolmogorov-Smirnov statistic

$$K_n=\sup_{\Theta\in\Xi}\sup_{z\in R_z}|T_n(\Theta,z)|$$

by an application of the continuous mapping theorem. In general the asymptotic distribution of K_n depends on certain features of the data generating process and bootstrap approximations are discussed in [3]. However, in the case where the pair (X,ε) and the covariate Z are independent the situation simplifies substantially. Here the covariance of the limiting process is given by (8) that

$$\mathrm{Cov}(\mathbb{T}(\Theta_1,y),\mathbb{T}(\Theta_2,z))=\tau(1-\tau)P(I\{X\in\Theta_1\cap\Theta_2\})(F_Z(y\wedge z)-F_Z(y)F_Z(z)),$$

where F_Z is the distribution function of the random variable Z and $y\wedge z$ denotes the vector of minima of the corresponding coordinates of y and z. If additionally X,Z are real-valued and $\Xi=\{(-\infty,t]|t\in\mathbb{R}\}$, the asymptotic covariance in Theorem 0.1 reduces to

$$\mathrm{Cov}(\mathbb{T}((-\infty,t],y),\mathbb{T}((-\infty,s],z))=\tau(1-\tau)F_X(s\wedge t)(F_Z(y\wedge z)-F_Z(y)F_Z(z)).$$

Hence, for univariate independent covariates X and Z with continuous distribution functions F_X and F_Z, respectively, the Kolmogorov-Smirnov test is asymptotically distribution-free because in this case the statistic

$$\sqrt{n}\sup_{x\in R_X,z\in R_Z}|T_n(x,z)|=\sqrt{n}\sup_{s,t\in[0,1]}|T_n(F_X^{-1}(s),F_Z^{-1}(t))|$$

converges in distribution to $\sqrt{\tau(1-\tau)}\sup_{s,t\in[0,1]}|B(s,t)|$, where B is the Kiefer-Müller process on $[0,1]^2$, i.e. a centered Gaussian process with covariance kernel

$$\mathrm{Cov}(B(s_1,t_1),B(s_2,t_2)) \quad = \quad (s_1\wedge s_2)(t_1\wedge t_2 - t_1t_2).$$

References

[1] H. Dette and S. Volgushev, *Non-crossing nonparametric estimates of quantile curves*, Journal of the Royal Statistical Society, Ser. B **70(3)** (2008), 609-627.

[2] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications*, Chapman & Hall (1996).

[3] S. Volgushev, M. Birke, H. Dette and N. Neumeyer, *Significance testing in quantile regression*, Electronic Journal of Statistics **7** (2013), 105-145.

[4] A.W. van der Vaart and J.A. Wellner, *Weak Convergence and Empirical Processes. Springer Series in Statistics*, Springer, New York (1996).

Higher dimensional quantiles and partial identification

Alfred Galichon

(joint work with Victor Chernozhukov and Marc Henry)

Consider a consumer choice problem (hedonic model) with quasilinear utility of the form:

$$\tilde{U}(x,z,\varepsilon) - p\,(z)$$

where, $x \in \mathbb{R}^{d_x}$ is a vector of observed consumer characteristics, $z \in \mathbb{R}^d$ is a vector of observed good characteristics. P_z denotes its distribution, and $P_{z|x}$ its distribution conditional on x.

$\varepsilon \in \mathbb{R}^d$ is a vector of unobserved consumer characteristics. P_ε will denote the distribution of ε, whether it is known a priori or not. P_ε has positive density on its domain, which is the closure of a connected open set. ε is assumed to be independent from x (this assumption may be relaxed as soon as the joint distribution of (x,ε) is fixed).

$p(z)$ is the observed price of good with characteristics z.

$\tilde{U}$ is an unknown utility function. Our focus is the identification of $\tilde{U}$, insofar as possible. We make the following separability assumption:

Assumption 1. *$\tilde{U}(x,z,\varepsilon) = U(x,z) + \zeta_\theta(x,z,\varepsilon)$, where ζ is a known function parameterized by θ.*

In the fist part of this talk we shall treat θ as fixed and we shall omit it.

One dimensional case. Here, we assume both z and ε have dimension one ($d = 1$) and we recall Matzkin's (2003) identification strategy. This strategy has been applied to the identification of hedonic models in Ekeland, Heckman and Nesheim (2004) and Heckman, Matzkin and Nesheim (2010).

We shall assume:

Assumption 2. *For every x, $\zeta(x,z,\varepsilon)$ is twice differentiable in z and ε, and*

$$\partial^2_{z\varepsilon}\zeta(x,z,\varepsilon) > 0.$$

Let $V(x,z) := p(z) - U(x,z)$. As p is known, identification of U is equivalent to identification of V.

Let $z(x,\varepsilon)$ be the quality purchased by consumer of type (x,ε), and let $\varepsilon(x,z)$ be its inverse w.r.t. its second variable. $\varepsilon(x,z)$ is implicitly defined by the first order conditions

$$-\partial_z V(x,z) + \partial_z \zeta(x,z,\varepsilon) = 0 \tag{1}$$

and we have:

Proposition 1. *Under the above assumptions, the map $z \to \varepsilon(x,z)$ is increasing,*

$$\varepsilon(x,z) = F^{-1}_{\varepsilon|X=x}\left(F_{Z|X=x}(z|x)\,|x\right)$$

and

$$z(x,\varepsilon) = F^{-1}_{Z|X=x}\left(F_{\varepsilon|X=x}(x,\varepsilon)\,|x\right).$$

We then recover V (and hence, U) by integration of the first order conditions (1)

$$V(x,z) = \int_0^z \partial_z\zeta(x,z',\varepsilon(x,z'))dz' + c.$$

Identification with a single market. Under Assumption 1 and recalling our notation $V(x,z) = p(z) - U(x,z)$, the consumer maximizes

$$\max_z \left(\zeta(x,z,\varepsilon) - V(x,z)\right).$$

The first order condition gives:

$$\nabla_z V(X,Z) = \nabla_z \zeta(X,Z,\varepsilon). \tag{2}$$

Identification of V is obtained via Optimal Transportation theory and the Monge-Kantorovich theorem (see Villani 2004, 2009). We need the following regularity conditions:

Assumption 3. *The following hold.*

1. ζ satisfies the Twist Condition relative to (z,ε): i.e.,

$$\nabla_z\zeta(x,z,\varepsilon_1) = \nabla_z\zeta(x,z,\varepsilon_2) \Rightarrow \varepsilon_1 = \varepsilon_2.$$

2. ζ is locally Lipschitz as a function of ε.

We have:

Theorem 1. *Under Assumptions 1 and 3, V is identified (up to a constant) as the solution to the variational problem*

$$\inf_V \left(\mathbb{E}[V(X,Z)] + \mathbb{E}[V^\zeta(X,\varepsilon)]\right),$$

where $V^\zeta(x,\varepsilon) = \sup_z\{\zeta(x,z,\varepsilon) - V(x,z)\}$.

This strategy was proposed in Galichon and Salanié (2012) in the context of identification of matching games, in the discrete case. It is also used in Chiong, Galichon and Shum (2013) for identification of dynamic discrete choice models. We can see this result as a multivariate generalization of the notion of quantile transform.

Parametric (partial) identification with multiple markets. The above analysis implies that for a given value of the parameter θ in $\zeta_\theta(x, z, \varepsilon)$, $U(x, z)$ is exactly identified provided the assumptions made hold. This implies that:
- the above analysis does not allow us to say anything about the estimation of θ
- if we observe multiple markets, θ will be overidentified.

Conversely, multiple market data may open up new possibilities for the estimation of θ.

Assumption 4. *The function ζ and the probability distribution P_ε are known up to a vector θ of unknown parameters and constant across markets.*

For the sake of simplicity, we shall consider the case of two markets. In each market, under Assumptions 1 and 3, we identify a function $U_m(x, z; \theta)$. Between markets m_1 and m_2, the distributions of producer and consumer characteristics may vary, hence the endogenous distribution of good characteristics and the price schedule. We assume that the utility function is unchanged and define the identified set Θ_I accordingly.

Proposition 2. *The identified set is equal to*

$$\Theta_I = \{\theta : U_{m_1}(x, z; \theta) = U_{m_2}(x, z; \theta),\ x, z \text{ - } a.s.\}.$$

We let $Q(\theta)$ be defined as $Q(\theta) = \mathbb{E}(U_{m_1}(x, z; \theta) - U_{m_2}(x, z; \theta))^2$. Based on estimators of U in each market, denoted $\hat{U}_m(x, z; \theta)$, we can base inference on the identified set on the quantity:

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{U}_{m_1}(x_i, z_i; \theta) - \hat{U}_{m_2}(x_i, y_i; \theta))^2$$

for some discretization $(x_i, z_i)_{i=1,\dots,n}$ of the space of characteristics.

References

[1] Chiong, K., Galichon, A. and Shum, M. (2013). "Estimating dynamic discrete choice models via convex analysis". Preprint.

[2] Ekeland, I., Galichon A. and Henry, M. (2012). "Comonotone measures of multivariate risks," *Mathematical Finance* 22 (1), pp. 109-132.

[3] Ekeland, I., Heckman, J., and Nesheim, L. (2004). "Identification and Estimation of Hedonic Models". *Journal of Political Economy* 112, pp. S60–S109.

[4] Galichon, A. and Henry, M. (2012). "Dual theory of choice with multivariate risks," *Journal of Economic Theory* 147, pp. 1501–1516.

[5] Galichon, A., Hallin, M., and Henry, M. (2013). "Monge-Kantorovich Quantiles and Multivariate Statistical Depth". In preparation.

[6] Galichon, A., and Salanié, B. (2013). "Cupid's invisible hand. Social surplus and identification in matching models". Preprint.
[7] Heckman, J., Matzkin, R., and Nesheim, L. (2010). "Nonparametric identification and estimation of nonadditive hedonic models". *Econometrica* 78, pp. 1569–1591
[8] Matzkin, R. (2003). "Nonparametric Estimation of Nonadditive Random Functions". *Econometrica* 71, pp. 1339–1375.
[9] Villani, C. (2003). *Topics in Optimal Transportation*, American Mathematical Society.
[10] Villani, C. (2009). *Optimal Transport. Old and New*. Springer.

Remarks on the EPK Puzzle. The (Im)Possibility of Demixing

Maria Grith
(joint work with Wolfgang K. Härdle, Ya'acov Ritov)

The motivation for this research topic are empirical findings on the financial markets that reveal a puzzling behavior of investors, which cannot be explained by means of traditional expected utility. The concept that will be used along this study is the pricing kernel (PK) that relates a continuous version of Arrow-Debreu pricing rule to the physical measure of the asset prices. Consistent with this, under risk neutral valuation in the arbitrage free models, the price at time t of the random payoff $\psi(S_T)$ is a martingale

$$V_t = \mathrm{E}^{Q_t}\left[\psi(S_T)\right] \tag{1}$$

$$= \mathrm{E}^{P_t}\left[\psi(S_T)\mathcal{K}_t(S_T)\right] \tag{2}$$

for $\{S_t\}_{t\in[0,T]}$ the price process of a risky asset with continuously distributed marginals and zero continuously compounded risk free interest rate; Q_t the conditional risk neutral measure. By Girsanov theorem we switched to pricing under the conditional physical measure P_t in (2). If the risk neutral and physical measures admit probability density functions q_t and p_t respectively, then

$$\mathcal{K}_t(s_T) = \frac{q_t(s_T)}{p_t(s_T)}, \tag{3}$$

for every realization s_T of S_T.

Preference based asset pricing models exploit the information embedded in the prices of financial assets to formally derive risk attitude parameters. They assume the existence of a representative agent whose marginal utility u' is proportional to the pricing kernel

$$\mathcal{K}_t(s_T) \propto u'(s_T)$$

Standard microeconomic theory assumes $u : \mathbb{R}_+ \to \mathbb{R}$ to be increasing, concave, twice continuously differentiable, hence the pricing kernel shall be *nonincreasing*. Starting with [1], [2], [3], different econometric methods have been applied to estimate the RHS of equation 3, with varying underlying models for the financial markets. It turned out as a common result, that typical estimates have nonmonotonic shape. This is what we call the *empirical pricing (EPK) kernel puzzle*.

Motivated by these findings, [5] provide an economic model that admits nonmonotone pricing kernels. They retain the expected utility framework in a single

period model and endow the financial investors with preferences that might be *state sensitive*. More technically, investors switch between two utility indexes at a point called *reference point*. As a consequence, while the utility indices are concave in individual wealth, the market utility may have jumps in the aggregate wealth space. In equilibrium, this may render pricing kernel nonmonotomic.

As an example, we consider $i = 1 \dots m$ investors that maximize a state dependent expected utility $\mathrm{E}^{P_t}\left[u^i\{S_T, y\}\right]$, subject to a budget constraint $B(y)$, with

$$u^i\{S_T, y\} = u^0\{y\}\,\mathbf{I}\{S_T \in [0, x_i]\} + u^1\{y\}\,\mathbf{I}\{S_T \in (x_i, \infty)\}$$

for $u^0(y) = b_0 u(y)$, $u^1(y) = b_1 u(y)$, $b_0, b_1 > 0$ and

$$u(y) = \begin{cases} \frac{y^{1-\gamma}}{1-\gamma} & \text{if } \gamma \neq 1 \\ \log(y) & \text{if } \gamma = 1. \end{cases}$$

x_i denote the reference points and $\gamma > 0$ the coefficient of relative risk aversion. Let $F(s_T)$ denote the cdf of the reference points

$$F(s_T) = m^{-1} \sum_{i=1}^{m} \mathbf{I}\{x_i \leq s_T\}.$$

Under some additional technical conditions, one can show that in equilibrium the pricing kernel has the following form

$$\mathcal{K}_{\theta,F}(x) = \left[\frac{x}{\{1 - F(x)\}\, b_0^{\frac{1}{\gamma}} + F(x)\, b_1^{\frac{1}{\gamma}}}\right]^{-\gamma} \tag{4}$$

for $\theta = (\gamma, b_0, b_1)^\top$. If $b_0 < b_1$, $\mathcal{K}_{\theta,F}(x)$ is not monotone in x; we will consider this case. Equation (4) may be rewritten as

$$x\mathcal{K}_{\theta,F}^{\frac{1}{\gamma}}(x) = b_0^{\frac{1}{\gamma}} + \left(b_1^{\frac{1}{\gamma}} - b_0^{\frac{1}{\gamma}}\right) F(x) \tag{5}$$

The only restrictions are the positivity of the parameter vector θ and F monotone non-decreasing, bounded between $[0, 1]$. Thus, $x\mathcal{K}_{\theta,F}^{1/\gamma}$ should be monotone non-decreasing bounded between $(b_0^{\frac{1}{\gamma}}, b_1^{\frac{1}{\gamma}})$. This yields the following restriction for γ:

$$\frac{1}{x} + \frac{1}{\gamma}\frac{\mathcal{K}'_{\theta,F}(x)}{\mathcal{K}_{\theta,F}(x)} \geq 0, \quad x \in (\alpha, \beta) \tag{6}$$

For any such γ we can solve

$$F(x) = \frac{x\mathcal{K}_{\theta,F}^{1/\gamma}(x) - b_0^{\frac{1}{\gamma}}}{b_1^{\frac{1}{\gamma}} - b_0^{\frac{1}{\gamma}}}.$$

One case in which γ can actually be identified is when either $F(\alpha_1) = 0$ for $\alpha_1 > \alpha$, or $F(\beta_1) = 1$ for $\beta_1 < \beta$. Then (6) is actually inequality on (α, α_1) and/or (β_1, β). Then

$$\gamma = \frac{x\mathcal{K}'_{\theta,F}(x)}{\mathcal{K}_{\theta,F}(x)}, \quad x \in (\alpha, \alpha_1) \text{ or } x \in (\beta_1, \beta).$$

[4] show that the intertemporal pricing kernel is inherently time varying and has some stable pattern. This justifies the choice of a smooth dynamic model for the pricing kernel

$$\mathcal{K}_{\theta_t,F_t}(x) = \left[\frac{x}{\{1 - F_t(x)\} b_{0t}^{\frac{1}{\gamma_t}} + F_t(x) b_{1t}^{\frac{1}{\gamma_t}}}\right]^{-\gamma_t}$$

with $\theta_t = (\gamma_t, b_{0t}, b_{1t})^\top$ and F_t cdf. We can use a scale/shift model for F_t

$$F_t(x) = F\left(\frac{x - a_t}{d_t}\right) \quad \text{for} \quad a_t \in \mathbb{R} \quad \text{and } d_t \in \mathbb{R}_+$$

and seemingly F is (partially) identifiable in this model. It is also possible to use state variables X_t to pin down $(\gamma_t, b_{0t}, b_{1t}, a_t, d_t)$ for parametric F.

We consider three situations in which this type of models shall be estimated.

Least squares. In practice, the pricing kernel is not observable and we will use some preestimate $\widehat{\mathcal{K}}_t$. If we assume that $y_{tj} = \widehat{\mathcal{K}}_t(s_j)$ for observation points s_j, $j = 1, \ldots, n$, is a sample of noisy curves s.t.

$$y_{tj} = \mathcal{K}_{\theta_t,F_t}(s_j) + \varepsilon_{tj} \quad \text{with } \varepsilon_{tj} \sim (0, \sigma_t^2),$$

the fitting problem involves finding

$$(\widehat{\theta}_t, \widehat{F}_t) = \arg\min_{\theta_t,F_t} \sum_{t=1}^{T} \sum_{j=1}^{n} \{y_{tj} - \mathcal{K}_{\theta_t,F_t}(s_j)\}^2.$$

Maximum likelihood. The physical density p_t can be recovered from the (known) risk neutral density q_t by means of PK

$$p_t(S_{t+1}|\theta_t, F_t) = \frac{\frac{q_t(S_{t+1})}{\mathcal{K}_{\theta_t,F_t}(S_{t+1})}}{\int \frac{q_t(x)}{\mathcal{K}_{\theta_t,F_t}(x)}}$$

Based on paired $(q_t, s_{t+1})_{t=0}^{T-1}$ we want to estimate

$$(\widehat{\theta}_t, \widehat{F}_t) = \arg\max_{\theta_t,F_t} \sum_{t=0}^{T-1} \log p_t(S_{t+1}|\theta_t, F_t).$$

Generalized method of moments. For a price vector $A_t = (A_{1t}, \ldots, A_{kt})^\top$ of k assets at time t, equation (2) reads as following

$$A_t = \mathrm{E}^{P_t}\left[\mathcal{K}_{\theta_t,F_t}(S_{t+1}) A_{t+1}\right].$$

For a cross section of asset prices $(A_t)_{t=1}^T$ we are interested in

$$(\widehat{\theta}_t, \widehat{F}_t) = \arg\min_{\theta_t, F_t} \left\{ g_T^\top(\theta_t, F_t) W^{-1} g_T(\theta_t, F_t) \right\}.$$

for some weighting matrix W and

$$g_T(\theta, F) = \sum_{t=0}^{T-1} \left\{ \mathcal{K}_{\theta_t, F_t}\left(S_{t+1}\right) A_{t+1}/A_t - 1_k \right\}.$$

References

[1] Y. Ait-Sahalia and A. W. Lo, *Nonparametric risk management and implied risk aversion*, Journal of Econometrics **94** (2000), 9–51.

[2] J. Jackwerth, *Recovering risk aversion from option prices and realized returns*, Review of Financial Studies **13** (2000), 433–451.

[3] R. F. Engle and J. V. Rosenberg,*Empirical pricing kernels*, Journal of Financial Economics **64** (2002), 341–372.

[4] M. Grith, W. K. Härdle and J. Park, *Shape Invariant Modelling Pricing Kernels and Risk Aversion*, Journal of Financial Econometrics **11** (2013), 370–399.

[5] M. Grith, V. Krätschmer and W. K. Härdle, *Reference Dependent Preferences and the EPK Puzzle*, Discussion Paper SFB 649 DP 2013-023, Submitted to Review of Finance on 10.04.13

Multidimensional statistical analysis of fMRI data in risk perception and investment decision study

Wolfgang Härdle

(joint work with Piotr Majer)

Decision making is a complex process of integrating and comparing various aspects of choice options. In the past years decision neuroscience has made important progress in grounding these aspects of decision making in neural systems, see [1]. Understanding which parts of the human brain are activated during decisions under risk and which neural processes underly (risky) investment decisions are important goals in neuroeconomics. Here, we analyze functional magnetic resonance imaging (fMRI) data on 17 subjects which were exposed to an investment decision task from [2].We obtain a time series of three-dimensional images of the blood-oxygen-level dependent (BOLD) fMRI signals.

Most of the fMRI studies used the general linear model (GLM). Though it has led to important insights into the neurobiological processes underlying cognition and emotion, the GLM approach has some important limitations. First, it focuses on task-related changes in the mean BOLD signal. Thereby, the GLM neglects information that might be carried by the variability of the BOLD signal, see [3]. Second, the GLM is a model-based approach, and can therefore only detect effects that were previously hypothesized and modeled. We apply a panel version of

the dynamic semiparametric factor model (DSFM) presented in [4] and identify task-related activations in space and dynamics in time.

$$Y^i_{t,j} = m_0(X_j) + \sum_{l=1}^{L} \left(\overline{Z}_{t,l} + \alpha^i_{t,l}\right)\ m_l(X_j) + \varepsilon_{t,j}, \tag{1}$$
$$1 \le j \le J,\ 1 \le t \le T,\ 1 \le i \le I\ .$$

Here, $Z_t = (\mathbf{1}, \overline{Z}_{t,1}, \ldots, \overline{Z}_{t,L})^\top$ is an unobservable $(L+1)$-dimensional stochastic process and m is an $(L+1)$-tuple $(m_0, \ldots, m_L)$ of unknown real-valued functions m_l. The voxel's index (i_1, i_2, i_3) is the covariate $X_{t,j}$ and the normalized BOLD signal of subject i is the dependent variable $Y^i_{t,j};\ j = 1, \ldots, J;\ t = 1, \ldots, T$. The errors $\varepsilon^i_{t,j}$ are assumed to be independent of $\overline{Z}_{t,j}$ and have zero means and finite second moments. The (common) functions m_l are approximated by a space basis $\Psi_{t,j} = [\psi_1(X_{t,j}), \ldots, \psi_K(X_{t,j})]^\top$ and corresponding $(L+1)\times K$ matrix of unknown coefficients A^*. More precisely, $[\psi_1(X_{t,j}), \ldots, \psi_K(X_{t,j})]^\top$ denote quadratic tensor B-splines on K equidistant knots. $\alpha^i_{t,l}$ is the fixed individual effect for subject i on function m_l at time point t. For identification purpose with respect to subjects, we assume that expectation of the individual effects over all subjects and over all functions m_l sums to zero, e.g.:

$$\mathrm{E}\left[\sum_{i=1}^{I}\left(\sum_{l=1}^{L}\alpha^i_{t,l} m_l(X_j)|X_j\right)\right] = 0. \tag{2}$$

With the panel DSFM (PDSFM) we can capture the dynamic behavior of the specific brain regions common for all subjects and represent the high-dimensional time series data in easily interpretable low dimensional dynamic factors without large loss of variability. After applying the PDSFM technique we estimated 20 spatial factors. 6 of them ($\widehat{m}_l\ , l = 5, 9, 12, 16, 17, 18$) correspond to brain areas medial Orbifrontal Cortex (mOFC) and Parietal Cortex (PC) which were already found in decision making contexts (see [1] for review). Beside these interesting factors connected with decision making, we detected other spatial maps that correspond to brain areas previously associated with motor responses and visual perception. These maps are likely unrelated to the decision making process within the task but confirming the activity of regions which were necessary to provide the answer by pushing a button.

The dynamics and subject specificity are jointly represented by the low-dimensional time series $\widehat{Z}^i_{t,l} = \overline{Z}_{t,l} + \alpha^i_{t,l}, i = 1, \ldots, I; l = 1, \ldots, L$. These subject-specific $\widehat{Z}^i_{t,l}$ correspond to the individual temporal differences of the activated brain regions in $\widehat{m}_l$. We find out that the responses to the stimulus of the weakly risk-averse individual show a significantly different volatility than the responses of the strongly risk averse individual. We found this volatility pattern in all factor loadings corresponding to the selected factors, e.g. for $l = 5, 9, 12, 16, 17, 18$. We classify studied subjects based on the standard deviation of the data extracted from the BOLD

signal, without knowing the subject's estimated risk attitude. Classification analysis of the subjects was conducted via Support Vector Machines (SVM), see [5]. Very high classification rates (97% for strongly and 75% for weakly risk-averse subjects) were obtained with the SVM classifier by applying the double cross validation algorithm. Herewith we have shown that our PDSFM approach is able to detect the neural representations of risk attitude and to classify the weak and strong averse individuals by their time-dependent factor loadings.

References

[1] H.R. Heekeren, S. Marrett, L.G. Ungerleider, *The neural systems that mediate human perceptual decision making*, Nature Reviews Neuroscience **9** (2008), 467–479.

[2] P. N. C. Mohr, G. Biele, L. K. Krugel, S. Li, H.R. Heekeren, *Neural foundations of risk-return trade-off in investment decisions*, NeuroImage **49** (2010), 2556–2563.

[3] P. N. C. Mohr, I. E. Nagel, *Variability in brain activity as an individual difference measure in neuroscience?*, The Journal of Neuroscience **30** (2010), 7755–7757.

[4] B. U. Park, E. Mammen, W. K. Härdle, S. Borak, *Time Series Modelling With Semiparametric Factor Dynamics*, Journal of the American Statistical Association **104** (2009), 284–298.

[5] C. Cortes, V. Vapnik, *The Nature of Statistical Learning Theory*, Machine Learning **20** (2005), 273–297.

Indentification and Shape Restrictions

Joel Horowitz

(joint work with Joachim Freyberger)

This talk is about estimation of the linear functional $L(g)$, where the unknown function g satisfies

$$Y = g(X) + U \tag{1}$$

and either

$$E(U|W = w) = 0 \tag{2}$$

or

$$P(U \leq 0|W = w) = q \tag{3}$$

for some q satisfying $0 < q < 1$ for almost every w.

Y is the dependent variable, X is a possibly endogenous explanatory variable, W is an instrument for X, and U is an unobservable random variable. The data consists of an independent random sample $\{Y_i, X_i, W_i : i = 1, ..., n\}$ from the distribution of (Y, X, W). It is assumed that X and W are discretely distributed random variables with finitely many mass points. Discretely distributed explanatory variables and instruments occur frequently in applied research. When X is discrete, g can be identified only at mass points of X. Linear functionals that may be of interest in this case are the value of g at a single mass point and the difference between the values of g at two different mass points. The model of equations (1) and (3) includes a class of nonseparable models.

In much applied research, W has fewer mass points than X does. The function g is not identified nonparametrically when W has fewer mass points that X. The linear functional $L(g)$ is unidentified except in special cases. Indeed, except in special cases, $L(g)$ can have any value in $(-\infty, \infty)$ when W has fewer points than X does. Thus, except in special cases, the data are uninformative about $L(g)$ in the absence of further information. In applied research, this problem is usually dealt with by assuming that g is a linear function. The assumption of linearity enables g and $L(g)$ to be identified, but it is problematic in other respects. In particular, the assumption of linearity is not testable if W is binary. Moreover, any other two-parameter specification is observationally equivalent to linearity and untestable, though it might yield substantive conclusions that are very different from those obtained under the assumption of linearity. For example, the assumptions that $g(x) = \beta_0 + \beta_1 x^2$ or $g(x) = \beta_0 + \beta_1 sinx$ for some constants β_0 and β_1 are observationally equivalent to $g(x) = \beta_0 + \beta_1 x$ if W is binary.

This talk explores the use of restrictions on the shape of g such as monotonicity, convexity, or concavity, to achieve partial identification of $L(g)$ when X and W are discretely distributed and W has fewer mass points than X has. Specifically, the talk uses shape restrictions on g to establish an identified interval that contains $L(g)$. Shape restrictions are less restrictive than a parametric specification such as linearity. They are often plausible in applications and may be prescribed by economic theory. For example, demand and cost functions are monotonic, and cost functions are convex. It is shown in this talk that under shape restrictions, such as monotonicity, convexity, or concavity, that impose linear inequality restrictions on the values of $g(x)$ at points of support of X, $L(g)$ is restricted to an interval whose upper and lower bounds can be obtained by solving mathematical programming problems. The estimated bounds are asymptotically distributed as the maxima of multivariate normal random variables. Under certain conditions, the bounds are asymptotically normally distributed, but calculation of the analytic asymptotic distribution is difficult in general. The bootstrap can be used to estimate the asymptotic distribution of the estimated bounds in applications. The asymptotic distribution can be used to carry out inference about the identified interval that contains $L(g)$ about the parameter $L(g)$.

Empirical Bayesian tuning parameters

Tatyana Krivobokova

Many penalized estimators have counterparts in the (empirical) Bayesian framework, with tuning parameters being an inverse scaling parameter of the prior distribution put on the mean of the data. Such tuning parameters are estimated from the corresponding likelihood and known to be sub-optimal in some models. At the same time, they are proved to be remarkably robust in praxis.
In this talk the empirical Bayesian tuning parameter for spline nonparametric estimators is discussed. This tuning parameter can be obtained assuming that the

underlying regression function is a realisation of a certain integrated Wiener process. It is well-known that such a tuning parameter is suboptimal with respect to L_2-risk. In particular, in contrast to an unbiased risk minimizing tuning parameter, the empirical Bayesian tuning parameter is not able to adapt to the unknown smoothness of the regression function.
Furthermore, estimators of empirical Bayesian and unbiased risk minimizing tuning parameters are studied. Their consistency and asymptotic normality are shown for the regression function from the Sobolev space of a given order. It is found that the convergence rate of tuning parameter estimators is very slow and agrees with known results from the kernel regression. Interesting insights deliver the obtained constants in the variances of both estimators. For the empirical Bayesian tuning parameter estimator the variance constant is found to be very small and fast decreasing with the penalty order parameter, while the variance constant of the unbiased risk minimizing tuning parameter estimator has opposite properties. Finally, it is discussed how the unknown smoothness of the regression function can be estimated from the data by comparing the estimating equations of both tuning parameters. The optimality of this procedure has not been studied yet.

References

[1] T. Krivobokova, *Smoothing parameter selection in two frameworks for penalized splines*, Journal of the Royal Statistical Society, Series B. In press.

Testing for a General Class of Functional Inequalities

Sokbae Lee
(joint work with Kyungchul Song, Yoon-Jae Whang)

This paper proposes a general testing method for inequality restrictions on nonparametric functions. More specifically, let $v_{\tau,1}, \cdots, v_{\tau,J}$ be nonparametric real-valued functions on $\mathbf{R}^d$ for each index $\tau \in \mathcal{T}$, where $\mathcal{T}$ is a subset of a finite dimensional space. This paper focuses on the problem of testing the following:

$$(1) \quad \begin{aligned} H_0 &: \max\{v_{\tau,1}(x), \cdots, v_{\tau,J}(x)\} \leq 0 \text{ for all } (x,\tau) \in \mathcal{X} \times \mathcal{T}, \text{ against} \\ H_1 &: \max\{v_{\tau,1}(x), \cdots, v_{\tau,J}(x)\} > 0 \text{ for some } (x,\tau) \in \mathcal{X} \times \mathcal{T}, \end{aligned}$$

where we take $\mathcal{X} \times \mathcal{T}$ to be a compact set.

This paper's framework is general, including many nonparametric testing problems in a unified framework. Among the examples are as follows:

(1) Testing inequality restrictions for conditional mean functions,
(2) Testing inequality restrictions for conditional quantile functions,
(3) Testing partial monotonicity of conditional distribution functions with respect to one of covariates, and
(4) Testing monotonicity of quantile regressions or interquartile functions.

Our test is easy to implement in general, mainly due to its recourse to the bootstrap method. The bootstrap procedure is based on a nonparametric bootstrap applied to kernel-based test statistics, while estimating "contact sets". This paper establishes the general asymptotic validity of the bootstrap procedure under high level conditions, and provide low level conditions for the examples listed above. Our bootstrap test is shown to exhibit good power properties. We also provide a general form of the local power function. In the paper, the asymptotic validity of the test is established uniformly over a large class of distributions. We support the usefulness of our testing approach by Monte Carlo experiments and applications to real-data examples.

Semi-parametric Bayesian Partially Identified Models

Yuan Liao
(joint work with Anna Simoni)

Partially identified models have been receiving extensive attentions in recent years, due to their broad applications in statistics, economics, education, engineering and many other fields in science and social science. Due to the limitation of the data generating process, the data cannot provide any information within the set where the structural parameter is partially identified (called *identified set*). One has to seek for "outside-data" information in order to explore more details inside the identified set. As a result, it should be desirable if a inference procedure can conveniently combine the information from both the observable data and other sources, i.e., economical theory, prior knowledge, experience, etc. A Bayesian approach is very appealing for partially identified models because it is convenient to take into account the subjective prior information, if any, so-called "outside-data information".

Bayesian analysis for partially identified models produces a posterior distribution that will asymptotically concentrate around the true identified set. When informative (subjective) priors are available for the structural parameter, the shape of the posterior density may not be flat even inside the identified set, providing more information about the parameter that cannot be told by the data. When no a priori information is available, using a uniform prior helps us estimate the true identified set. Therefore, the asymptotic behavior for the posterior distribution is different from that of the traditional point identified case; the latter is usually normally distributed due to the Bernstein von Mises theorem, and hence the information from the prior is often washed away by the data when the structural parameter is identifiable.

1. Two existing Bayesian approaches

There are in general two Bayesian approaches for partially identified models. The first one is based on a known parametric likelihood function (e.g., Moon and Schorfheide 2012, Poirier 1998), which also involves a a finite dimensional nuisance parameter, denoted by ϕ. Then the identified set is completely determined by ϕ,

denoted by $\Theta(\phi)$. Besides known likelihood functions, another important feature of this approach is that the prior for θ is imposed conditional on ϕ only (e.g., uniform on $\Theta(\phi)$), hence it must incorporate the partial identification structure. The limitation of this approach is that it requires an ad-hoc parametric form of the likelihood. Econometric models, on the other hand, often only identify a set of "moment inequalities" instead of a known likelihood function. Therefore once the function form is incorrectly specified, the posterior can be misleading. Hence robustness is a big question.

The second approach only requires a set of moment inequalities, and uses a moment-condition-based likelihood (Liao and Jiang 2010). By doing so it successfully avoids assuming the knowledge of the true likelihood function. In addition, the prior $\pi(\theta)$ can be placed marginally (e.g., $N(0,1)$), hence the prior does not need to take into account the partial identification restriction. However, this approach does not possess a pure probabilistic interpretation, which only uses a Bayesian machinery to make quasi-Bayesian inference. How close the calculated posterior is to the true posterior is largely unknown.

2. A new semi-parametric Bayesian procedure

We propose a semi-parametric Bayesian procedure for inference. The proposed procedure is pure Bayesian, so it has a well defined probabilistic interpretation. More importantly, it does not require a known parametric form of the likelihood function, but only a set of moment conditions. Therefore it solves the robustness issue. We place a prior $\pi(l)$ on the unknown likelihood function; the latter can be either a CDF $l = F$, or a density $l = f$. The unknown likelihood can be written as $l(D_n, \phi)$, where ϕ is point identified but nuisance. The parameter of interest is θ, and $D_n = \{X_i\}_{i=1}^n$ are observed data. When $l = F$, we specify a Dirichlet process prior $\pi(F)$ which then deduces a prior on ϕ through $\phi(F)$. The Bayesian experiment is

$$X|F \sim F, \qquad F \sim \pi(F) : DirichletProcess, \qquad \theta|\phi = \phi(F) \sim \pi(\theta|\phi(F))$$

Other priors such as the Polya tree can be used for $\pi(F)$ too. Let $p(F|D_n)$ denote the marginal posterior of F, given by $p(F|D_n) \propto \pi(F)\prod_{i=1}^n F(X_i)$. Then the marginal posterior density function of θ writes

$$p(\theta|D_n) = \int p(\theta|\phi(F), D_n)p(F|D_n)dF = \int_{\mathcal{F}} \pi(\theta|\phi(F))p(F|D_n)dF. \tag{1}$$

3. Asymptotic behavior

Let us assume there is a true value of ϕ, denoted by ϕ_0, which induces a true identified set $\Theta(\phi_0)$. Define the ϵ-envelope of a set $\Theta(\phi)$ as $\Theta(\phi)^\epsilon = \{\theta : d(\theta, \Theta(\phi)) \leq \epsilon\}$ where $d(\theta, \Theta(\phi)) = \inf_{x\in\Theta(\phi)} \|\theta - x\|$. We achieve the *posterior consistency* for partial identification: for any $\epsilon > 0$,

$$P(\theta \in \Theta(\phi_0)^\epsilon | D_n) \to^p 1.$$

The large sample property, such as the posterior consistency and concentration rate, is one of the benchmarks of a Bayesian procedure under consideration, which ensures that with a sufficiently large amount of data, it is nearly possible to recover the truth identified set. Therefore lack of consistency is extremely undesirable.

For the true identified set $\Theta(\phi_0)$, the posterior concentration rate is described under the Hausdorff distance: for some $C > 0$,

$$P(d_H(\Theta(\phi), \Theta(\phi_0)) \leq C\sqrt{\frac{\log n}{n}}|D_n) \to^p 1.$$

4. Bayesian Analysis for Support Function

In partially identified models, the identified set $\Theta(\phi)$ becomes one of the important objects to study and to make inference. When $\Theta(\phi)$ is convex, the support function provides us a convenient way to characterize the identified set, which is defined as

$$S_\phi(p) = \sup_{\theta \in \Theta(\phi)} \theta^T p$$

where $p \in \mathbb{S}^{\dim(\theta)}$, the unit sphere. Any non-empty closed convex set is uniquely determined by its support function. For example, θ is inside the closure $\overline{\Theta(\phi)}$ if and only if for all $\|p\| = 1$, $\theta^T p = S_\phi(p)$. A a result, the support function has been a useful tool for analyzing partially identified models, e.g., Beresteanu and Molinari (2008) and Kaido and Santos (2012), Chandrasekhar et al. (2012), Bontemps et al. (2010), etc.

The posterior of $S_\phi(\cdot)$ thus is very useful for inference about the identified set, which is determined by that of ϕ. By putting a prior on $S_\phi(\cdot)$ via the prior on ϕ, we can obtain the posterior. When $S_\phi(\cdot)$ is treated as an operator of ϕ, it can be highly nonlinear, which is hard to deal with. We derive a local linear approximation to the support function: There is a vector $A(p, \phi_0)$ that depends on p and ϕ_0 only, such that for the ball $B(\phi_0, \frac{C}{\sqrt{n}})$,

$$\sup_{\phi_1, \phi_2 \in B(\phi_0, Cn^{-1/2})} \sup_{\|p\|=1} \sqrt{n}|(S_{\phi_1}(p) - S_{\phi_2}(p)) - A(p, \phi_0)^T(\phi_1 - \phi_2)| = o(1), \tag{2}$$

for some $C > 0$. Equation (2) then implies the Bernstein von Mises theorem for the support function, which is, the posterior of $\sqrt{n}(S_\phi(p) - S_{\hat{\phi}}(p))$ is asymptotically normal for each p, where $\hat{\phi}$ is the posterior mode of $p(\phi|D_n)$. This is the semi-parametric BvM theorem for $S_\phi(\cdot)$, which provides us a useful way to approximate the posterior of the support function under large sample.

5. Optimality based on Bayesian Decision Making

Unlike the classical identifiable case, the optimality of set estimation for $\Theta(\phi)$ is not well studied in the literature, partially due to the lack of proper assessment for loss. A few important contributions are done from the inference and frequentist perspective, (e.g., Canay 2010, Kaido and Santos 2012, etc.)

Optimality for estimating the identified set can be achieved based on Bayesian decision making. For any estimator Ω of the identified set, let

$$\Theta(\phi)\Delta\Omega = (\Theta(\phi) \cap \Omega^c) \cup (\Theta(\phi)^c \cap \Omega)$$

be the symmetric difference between $\Theta(\phi)$ and Ω, and $\mu(\Theta(\phi)\Delta\Omega)$ be its Lebesgue measure. Define the Bayesian loss function

$$L(\Omega) \equiv E[\mu(\Theta(\phi)\Delta\Omega)|D_n],$$

where the expectation is taken with respect to the posterior of $\Theta(\phi)$. The following optimality of estimating the identified set can be shown:

$$\widehat{\Omega} \equiv \{x \in \Theta : P(x \in \Theta(\phi)|D_n) \geq 0.5\} = \arg\min_{\Omega} L(\Omega). \tag{3}$$

where $P(x \in \Theta(\phi)|D_n)$ is a probability measure taken with respect to the posterior of $\Theta(\phi)$ for a fixed x in the parameter space of θ. Therefore $\widehat{\Omega}$ is optimal in the Bayesian-decision-making sense as it minimizes the Bayesian risk.

6. Bayesian Credible Set

Bayesian inference can be carried out through finite-sample Bayesian credible sets (BCS), which is a set BCS(τ) such that

$$P(\theta \in \text{BCS}(\tau)|D_n) = 1 - \tau$$

at level $1 - \tau$. Things become more interesting when we construct the BCS for $\Theta(\phi)$. Based on the support function, we can construct two-sided credible sets $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ such that

$$P(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}|D_n) = 1 - \tau, \tag{4}$$

which gives the two-sided BCS for the identified set. Here q_τ is some quantile derived based on the posterior of the support function, and $\widehat{\phi}$ is still the posterior mode for ϕ.

The BCS for θ does not have a correct frequentist coverage when θ is partially identified, since the BCS tends to be a subset of the interior of the frequentist confidence set, as shown by Moon and Schorfheide (2012) when the parametric likelihood is known. In addition, Gustafson (2012) showed that from a frequntist point of view, there is always a region inside the identified set which Bayesian credible interval fails to cover.

In contrast, the BCS for the identified set has desired frequentist coverages. Specifically, for any $\tau > 0$,

$$P(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}) \geq 1 - \tau + o_p(1). \tag{5}$$

The rationale behind (5) is that, the identified set itself is "point identified", whose posterior will concentration around a neighborhood of the true set. Hence the support of the posterior of the set is always larger than the set. Since the Bernstein von Mises theorem now holds on the identified set (through its support function), the posterior BCS has the correct frequentist coverage probability.

7. About the authors

Both authors have strong research interests in theoretical and applied Bayesian econometrics. In addition, the authors are also interested in high-dimensional sparse modeling, factor analysis and inverse problems.

References

[1] Beresteanu, A. and Molinari, F. (2008) Asymptotic properties for a class of partially identified models. *Econometrica*, **76**, 763-814.

[2] Bontemps, C., Magnac, T. and Maurin, E. (2011). Set identified linear models. *Econometrica*, forthcoming.

[3] Canay, I. (2010) EL Inference for partially identified models: large deviations optimality and bootstrap validity. *Journal of Econometrics*. **156**, 408-425.

[4] Chandrasekhar, A., Chernozhukov, V., Molinari, F. and Schrimpf, P. (2012) Inference for best linear approximations to set identified functions. *Manuscript*. MIT.

[5] Gustafson, P. (2012) On the behaviour of Bayesian credible intervals in partially identified models. *Electronic Journal of Statistics*. **6**, 2107-2124.

[6] Kaido, H. and Santos, A. (2011). Asymptotically efficient estimation of models defined by convex moment inequalities, preprint.

[7] Liao, Y. and Jiang, W. (2010) Bayesian analysis in moment inequality models. *Annals of Statistics*. **38**, 275-316.

[8] Moon, H. R. and Schorfheide, F. (2012). Bayesian and frequentist inference in partially-identified models. *Econometrica*, **80**, 755-782.

[9] Poirier, D. (1998) Revising beliefs in nonidentified models. *Econometric Theory*. **14**, 483-509.

Nonparametric Tests for Regression Quantiles

Enno Mammen

(joint work with Ingrid van Keilegom, Kyusang Yu)

Consider a data set of n i.i.d. tuples (X_i, Y_i) where Y_i is a one-dimensional response variable, X_i is a d-dimensional covariate. For $0 < \alpha < 1$ we denote the conditional α-quantile of Y_i given $X_i = x$ by $r_\alpha(x)$. Thus we can write

$$Y_i = r_\alpha(X_i) + \varepsilon_{i,\alpha} \qquad (i = 1, \ldots, n), \tag{1}$$

with error variables $\varepsilon_{i,\alpha}$ that fullfill $q_\alpha(\varepsilon_{i,\alpha}|X_i) = 0$. Here, $q_\alpha(\varepsilon_{i,\alpha}|X_i)$ is the α-quantile of the conditional distribution of $\varepsilon_{i,\alpha}$ given X_i. A kernel estimator $\widehat{r}_\alpha$ of the regression quantile r_α is given by:

$$\widehat{r}_\alpha(x) = \arg\min_r \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \tau_\alpha(Y_i - r),$$

where $\tau_\alpha(u) = \alpha u_+ - (1-\alpha)u_-$ with $u_+ = uI(u > 0)$ and $u_- = uI(u < 0)$, is the check function and $K(u_1, \ldots, u_d) = \prod_{j=1}^d k(u_j)$ is a multivariate product kernel with one-dimensional density functions k defined on $[-1, 1]$ as factors and d-dimensional bandwidth parameter $h = (h_1, \ldots, h_d)$. For simplicity of notation we assume $h_1 = \ldots = h_d$ and we write also $h = h_j$.

The classical mathematical approach for the asymptotic analysis of quantile estimators is based on Bahadur expansions. The Bahadur expansion $\widetilde{r}_\alpha(x)$ of the kernel estimator $\widehat{r}_\alpha(x)$ is given by

$$\widetilde{r}_\alpha(x) \quad = \quad -\frac{\sum_{i=1}^n K\Big(\frac{x-X_i}{h}\Big)\{I(\varepsilon_{i,\alpha} \le 0) - \alpha\}}{\sum_{i=1}^n K\Big(\frac{x-X_i}{h}\Big) f_{\varepsilon_\alpha|X}(0|X_i)},$$

where $f_{\varepsilon_\alpha|X}(0|x)$ is the conditional density of ε_α given $X = x$. For the approximation error of the Bahadur expansion one can show that

$$\sup_{\alpha\in A}\sup_{x\in R_X}\Big|\widehat{r}_\alpha(x) - \widetilde{r}_\alpha(x)\Big| = O_P((nh^d)^{-3/4}L_n),$$

where L_n is a sequence that is of order $O((\log n)^C)$ for some $C > 0$, where the interval $A = [a, b]$ is a subset of $(0, 1)$, and where R_X is the support of X. For recent discussions of Bahadur expansions in nonparametrics see [1], [2], [3], and [4].

In this note we will discuss examples where the accuracy of the Bahadur expansion does not suffice for an asymptotic analysis. Our first example that we only shortly discuss comes from semiparametrics and it is the estimation of linear functionals

$$\int_{R_X} w(x)r_\alpha(x)dx.$$

Here one needs that the approximation achieves faster rates than the parametric rate $n^{-1/2}$:

$$\int_{R_X} w(x)(\widehat{r}_\alpha(x) - \widetilde{r}_\alpha(x))dx = o_P(n^{-1/2}).$$

This requires $(nh^d)^{-3/4} = o(n^{-1/2})$ and thus $n^{-1/3} \ll h^d$. In particular, it excludes the case that h^d is of order $n^{-1/2}$. For that case $\widetilde{r}_\alpha(x) - r_{0,\alpha}(x)$ is of order $n^{-1/4}$ and the stochastic behavior of $\int_{R_X} w(x)\widetilde{r}_\alpha(x)dx$ changes, see [5] for a discussion of this issue and for the theory of higher order inference functions. In [6] the case $h^d \sim n^{-1/2}$ is analyzed for average derivative estimation. For related discussions on quantile regression we see that approximations based on Bahadur expansions are too crude to allow such an asymptotic study.

We now come to another example where the accuracy of the Bahadur expansion is too weak. This is the asymptotic analysis of nonparametric tests of the hypothesis: $r_\alpha(x) = r_{0,\alpha}(x)$ for $\alpha \in A$, $x \in R_X$. Here, $r_{0,\alpha}$ is some specified function. In a more general set-up that will be used below it is a parametrically estimated function. We consider the following test statistic:

$$\int_A \int_{R_X} w(x, \alpha)(\widehat{r}_\alpha(x) - r_{0,\alpha}(x))^2 dx d\alpha$$

with some weight function $w(x, \alpha)$ or

$$\int_{R_X} w(x)(\widehat{r}_\alpha(x) - r_{0,\alpha}(x))^2 dx$$

with some weight function $w(x)$ if $A = \{\alpha\}$.

An analysis based on Bahadur expansions needs here that

$$\int_A \int_{R_X} w(x,\alpha) \left[(\widehat{r}_\alpha(x) - r_{0,\alpha}(x))^2 - (\widetilde{r}_\alpha(x) - r_{0,\alpha}(x))^2 \right] dx d\alpha$$

is of lower order, or that

$$\int_{R_X} w(x) \left[(\widehat{r}_\alpha(x) - r_{0,\alpha}(x))^2 - (\widetilde{r}_\alpha(x) - r_{0,\alpha}(x))^2 \right] dx$$

is of lower order, respectively. One can check that this is the case if $(nh^d)^{-5/4} = o((nh^d)^{-1}h^{d/2})$, or equivalently, if $n^{-1/3} \ll h^d$. For twice differentiable functions this allows to choose bandwidths that are rate optimal for testing or for estimation, respectively, only for the one-dimensional case $d = 1$. Thus again, a direct application of Bahadur expansions excludes interesting cases. For this reason, we will use a more refined approach.

We will consider the following testing problem:

$$H_0 : \text{For all } \alpha \in A \text{ there exists a } \theta(\alpha) \in \Theta, \text{ such that : } r_\alpha = r_{\alpha,\theta(\alpha)},$$

where $r_{\alpha,\theta}$ is a parametric family for all $\alpha \in A$. We assume that there are parametric estimators $\widehat{\theta}(\alpha)$ of α. We use the following test statistic.

$$\widehat{T}_A = \int_A \int \widehat{m}_\alpha^2(x) w(x,\alpha) dx d\alpha, \tag{2}$$

for some weight function $w(x,\alpha)$. For the case that A contains only one value α we use

$$\widehat{T}_\alpha = \int_{\mathcal{X}} \widehat{m}_\alpha^2(x) w(x) dx, \tag{3}$$

where

$$\widehat{m}_\alpha(x) = \arg\min_r \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \tau_\alpha(Y_i - r_{\alpha,\widehat{\theta}(\alpha)}(X_i) - r).$$

This test statistic is related to similar tests in mean regression for i.i.d. and time series data, see [7], [8], [9], [10], [11] and [12]. We assume for the quantile regression function:

$$r_\alpha(x) = r_{\alpha,\theta_0(\alpha)}(x) + n^{-1/2}h^{-d/4}\Delta_\alpha(x). \tag{4}$$

For the case $\Delta_\alpha \equiv 0$ the function r_α lies on the hypothesis. For the bandwidths we assume that $nh^{3d/2} \to \infty$. We have the following results.

THEOREM 1. *It holds that*

$$nh^{d/2}\widehat{T}_\alpha - b_{h,\alpha} \xrightarrow{d} N(D_\alpha, V_\alpha),$$

where

$$\begin{aligned} D_\alpha &= \int_{R_X} \Delta_\alpha(x)^2 w(x)\, dx, \\ b_{h,\alpha} &= h^{-d/2} K^{(2)}(0)\alpha(1-\alpha) \int_{R_X} \frac{w(x)}{f_X(x) f^2_{\varepsilon_\alpha|X}(0|x)} dx, \\ V_\alpha &= 4K^{(4)}(0)\alpha^2(1-\alpha)^2 \int_{R_X} \frac{w^2(x,\alpha)}{f^2_X(x) f^4_{\varepsilon_\alpha|X}(0|x)} dx. \end{aligned}$$

THEOREM 2. *It holds that*

$$nh^{d/2}\widehat{T}_A - b_{h,A} \xrightarrow{d} N(D_A, V_A),$$

where

$$\begin{aligned} D_A &= \int_A \int_{R_X} \Delta_\alpha(x)^2 w(x,\alpha)\, dx\, d\alpha, \\ b_{h,A} &= h^{-d/2} K^{(2)}(0) \int_A \alpha(1-\alpha) \int_{R_X} \frac{w(x,\alpha)}{f_X(x) f^2_{\varepsilon_\alpha|X}(0|x)} dx\, d\alpha, \\ V_A &= 4K^{(4)}(0) \int_{\alpha,\beta\in A, \alpha<\beta} \alpha^2(1-\beta)^2 \int_{R_X} \frac{w^2(x,\alpha)}{f^2_X(x) f^4_{\varepsilon_\alpha|X}(0|x)} dx\, d\alpha\, d\beta. \end{aligned}$$

Here is a short outline of the proof. Assume for simplicity that $A = \{\alpha\}$, $r_\alpha(x) \equiv 0$, $\Delta_\alpha(x) \equiv 0$, $r_{\alpha,\theta} \equiv r_\alpha$, and $w(x) \equiv 1$. Then we have that

$$\widehat{m}_\alpha(x) = \arg\min_r \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \tau_\alpha(\varepsilon_{i,\alpha} - r),$$

$$\widetilde{m}_\alpha(x) = -\frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\{I(\varepsilon_{i,\alpha} \le 0) - \alpha\}}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) f_{\varepsilon_\alpha|X}(0|X_i)},$$

We have to show that $\Delta_n = \int \widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2\, dx = o_P((nh^d)^{-1}h^{d/2})$. We write $\Delta_n = \Delta_{n,1} + \Delta_{n,2}$, where

$$\begin{aligned} \Delta_{n,1} &= \int [\widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2] - E^*[\widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2]\, dx \\ \Delta_{n,2} &= \int E^*[\widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2]\, dx \end{aligned}$$

with some suitable conditional expectation E^*.

For the treatment of $\Delta_{n,1}$ we use that $\widetilde{m}_\alpha(x_1)^2 - \widehat{m}_\alpha(x_1)^2$ and $\widetilde{m}_\alpha(x_2)^2 - \widehat{m}_\alpha(x_2)^2$ are independent for $||x_1 - x_2|| > h$. Thus

$$\begin{aligned}\Delta_{n,1} &= \int [\widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2] - E^*[\widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2]\, dx \\ &= \int [\widetilde{m}_\alpha(x) - \widehat{m}_\alpha(x)][\widetilde{m}_\alpha(x) + \widehat{m}_\alpha(x)] - E^*[\widetilde{m}_\alpha(x)^2 - \widehat{m}_\alpha(x)^2]\, dx \\ &= O_P(L_n(nh^d)^{-3/4}(nh^d)^{-1/2}h^{d/2}).\end{aligned}$$

For the treatment of $\Delta_{n,2}$ note that

$$\widehat{m}_\alpha(x) \le u \quad \text{if and only if} \quad \sum K\Big(\frac{x - X_i}{h}\Big)\Big\{I(\varepsilon_{i,\alpha} \le u) - \alpha\Big\} \ge 0.$$

The essential idea here is to use Edgeworth expansions of the right hand side to get expansions of $E^*[\widehat{m}_\alpha(x)^2]$.

References

[1] A. El Ghouch, I. Van Keilegom, I. *Local linear quantile regression with dependent data.* Statistica Sinica **19** (2009), 1621–1640.

[2] E. Guerre, C. Sabbah. *Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function.* Econometric Theory **28** (2012) 87–129.

[3] S. Hoderlein, E. Mammen. *Identification and estimation of local average derivatives in non-separable models without monotonicity.* Econometrics Journal **12** (2009) 1–25.

[4] E. Kong, O. Linton, and Y. Xia . *Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model.* Econometric Theory **26** (2010) 1529–1564.

[5] J. Robins, L. Li, E. Tchetgen,A. van der Vaart, *Higher order influence functions and minimax estimation of nonlinear functionals.* In Probability and statistics: essays in honor of David A. Freed- man. Inst. Math. Stat. Collect., Vol. 2. Inst. Math. Statist., (2008), Beachwood, OH, 335-421.

[6] M. Cattaneo, M.; R.K. Crump, M. Jansson *Generalized jackknife estimators of weighted average derivatives*, J. Amer. Statist. Assoc. (2013), to appear.

[7] Y. Ait-Sahalia, J. Fan, H. Peng. *Nonparametric transition-based tests for diffusions.* Journal of American Statistical Association **104** (2009) 1102–1116.

[8] H. Dette, I. Sprekelsen. *Some comments on specification tests in nonparametric absolutely regular processes.* J. Time Series Anal. **25** (2004) 159–172.

[9] J. Fan,C. Zhang, J. Zhang. *Generalized likelihood ratio statistics and Wilks phenomenon.* Annals of Statistics **29** (2001) 153–193.

[10] W. Härdle, E. Mammen. *Testing parametric versus nonparametric regression.* Ann. Statist. **21** (1993) 1926–1947.

[11] J.P. Kreiss, M.H. Neumann, Q. Yao. *Bootstrap tests for simple structures in nonparametric time series regression.* Statistics and its interface **1** (2008) 367–380.

[12] A. Leucht. *Degenerate U- and V-statistics under weak dependence: Asymptotic theory and bootstrap consistency.* Bernoulli **18** (2012) 552–585.

Min-wise hashing for large-scale regression analysis. Computational limits of identifiability

Nicolai Meinshausen
(joint work with Rajen Shah)

We study large-scale regression analysis in a "large p, large n" context for a linear regression model $Y = X\beta^* + \delta + \varepsilon$, where $Y \in \mathbb{R}^n$ is the response, $X \in \{0,1\}^{n\times p}$ a sparse binary predictor matrix, $\beta^* \in \mathbb{R}^n$ an optimal regression vector and $\delta, \varepsilon \in \mathbb{R}^n$ are the structural error and the independent noise term. While we have to make sparsity assumptions on β^* in the high-dimensional setting of "large p, small n" settings, no such assumptions are typically required for large-scale regression analysis where the number of observations n can (but does not have to) exceed the number of variables p. The main difficulty is that computing an OLS or ridge-type estimator is computationally infeasible for $n, p > 10^5$ and we need to find computationally efficient ways to approximate these solutions without increasing the prediction error by a large amount. Both n and p are often in the millions or larger for many recent applications such as text analysis, drug safety studies and web-scale prediction tasks. Trying to find interactions amongst millions of variables seems to be an even more daunting task. We study a small variation of the b-bit minwse-hashing scheme (Li and Konig, 2011) and show that the regression problem can be solved in a much lower-dimensional setting as long as $q\|\beta^*\|_2^2/n \to 0$ for $n \to \infty$, where q is the average number of non-zero entries in each row of the predictor matrix. We get finite-sample bounds on the prediction error. The min-wise hashing scheme is also shown to fit interaction models. Fitting interactions does not require an adjustment to the method used to approximate linear models, it just requires a higher-dimensional projection. We show some examples for simulated data and an application to detection of malicious URLs.

References

[1] P. Li and A.C. Konig, *Theory and applications of b-bit minwise hashing*, Communications of the ACM **54** (2011), 101–109.

Selected topics on selections of random sets

Ilya Molchanov

The partially identified objects can be represented as random sets, whereas the full identification setting corresponds to a random singleton viewed as a selection of this random set. This relation between random sets and random elements is similar to the relationship between subadditive set functions (capacities or non-additive measures) on one hand and σ-additive functions (probability measures) on the other one.

The two main settings typical for partially identified problems are

- The distribution (theoretical or empirical) of random set X is known and one is looking for its selections, possibly satisfying some extra properties.

- Selections of X are observed and the goal is to make inference about the distribution of X.

In any case the relationship between selections and random sets plays the crucial rôle. In the following we consider only random closed sets in the Euclidean space $\mathbb{R}^d$, see [7] for a comprehensive presentation of the theory of random sets.

Recall that random vector ξ is called a selection of random set X if $\xi \in X$ a.s. One often speaks about the ordered coupling of ξ and X. The probability measure μ is a distribution of a selection of a random closed set X if and only if

$$\mu(K) \le T(K) = \mathbf{P}\{X \cap K \neq \emptyset\} \tag{1}$$

for all compact sets K, see [1]. Equivalently,

$$\mathbf{P}\{\xi \in F\} \ge \mathbf{P}\{X \subset F\}$$

for all closed sets F. It is well known that each a.s. non-empty random closed set admits a selection and the set can be represented as the convex hull of a countable family of its selections. Furthermore, a characterisation of subsets of $L^p(\mathbb{R}^d)$ that can be interpreted as selections of a random closed set is available (for $p \in [1, \infty]$). A characterisation in the case of $p \in [0, 1)$ is still unknown.

The family of all compact sets in (1) can be replaced by the so-called core-determining class $\mathcal{M}$ [6]. A core determining class is distribution determining, but not the other way, e.g. if X is a convex compact random set, then the family of convex compact sets is distribution determining but not core determining.

Let w be a covariate. Then (ξ, w) can be realised as a selection of $X \times \{w\}$ if and only if

$$\mathbf{P}\{\xi \in K|\mathfrak{B}\} \le \mathbf{P}\{X \cap K \neq \emptyset|\mathfrak{B}\}\,,$$

where $\mathfrak{B}$ is generated by w. In particular, μ is the distribution of a selection independent of w if and only if

$$\mu(K) \le \operatorname{essinf} \mathbf{P}\{X \cap K \neq \emptyset|\mathfrak{B}\}\,.$$

see [2] in relation to treatment response. The talk further discusses the selections of set-valued processes and relationships to the no-arbitrage problem in proportional transaction costs models, see [8].

Taking the mean of all integrable selections yields the selection expectation of a random closed set. The advantage of using the selection expectation lies in its rather easy computation and has been used in [3] in relation to identification of the model from the observed equilibria in games with mixed strategies. The key argument is that the selection expectation is a convex set $\mathbf{E}X$ whose support function equals the expected support function of the random set X.

The talk addresses the existence issues for selections with given moments. For instance, let X be a random subset of the line. Then it possesses a selection ξ with $\mathbf{E}\xi = m_1$ and $\mathbf{E}\xi^2 = m_2$ if and only if the auxiliary random set

$$Y = \{(x, x^2) : \ x \in X\}$$

satisfies $\mathbf{E}Y \ni (m_1, m_2)$. The idea of auxiliary random sets appeared first in [4]. This can be generalised for the infinite moment sequence or values of characteristic functions and enables to determine if X possesses, e.g. a Gaussian selection.

Along the same line, it is possible to find the selection $\xi = (\xi_1, \xi_2)$ of square integrable random set $X \subset \mathbb{R}^2$ with maximum correlation between ξ_1 and ξ_2. For this, create an auxiliary random set

$$Y = \{(x_1, x_2, x_1^2, x_2^2, x_1 x_2) : \ (x_1, x_2) \in X\} \subset \mathbb{R}^5$$

and maximise

$$\frac{y_5 - y_1 y_2}{\sqrt{y_3 - y_1^2}\sqrt{y_4 - y_2^2}}, \qquad (y_1, \dots, y_5) \in \mathbf{E}Y .$$

Exploring all selections of a random set is also important in view of applications to multivariate risk measures. Let X be a random set that represents all possible portfolios that may be realised after admissible transactions at a terminal time. Then X is called acceptable if it possesses a selection with all individually acceptable (under certain law invariant coherent risk measures) marginals, see [5]. It is shown that the obtained set-valued risk measures admit a dual-representation akin to the classical case of univariate coherent risk measures.

References

[1] Z. Artstein. *Distributions of random sets and random selections* Israel J. Math. **46** (1983), 313–324.

[2] A. Beresteanu, I. Molchanov, and F. Molinari. *Partial identification using random sets theory* J. of Econometrics **166** (2011), 17–32

[3] A. Beresteanu, I. Molchanov, and F. Molinari. *Sharp identification regions in models with convex moment predictions* Econometrica **79** (2011), 1785–1821.

[4] A. Beresteanu and F. Molinari. *Asymptotic properties for a class of partially identified models* Econometrica **76** (2008), 763–814.

[5] I. Cascos and I. Molchanov. *Multivariate risk measures: a constructive approach based on selections* Technical report, Arxiv Math 1301:1496, 2013.

[6] A. Galichon and M. Henry. *Set identification in models with multiple equilibria* Review of Economic Studies **78** (2011), 1264–1298.

[7] I. Molchanov. *Theory of Random Sets*. Springer, London, 2005.

[8] W. Schachermayer. *The fundamental theorem of asset pricing under proportional transaction costs in finite discrete time* Math. Finance **14** (2004), 19–48.

Computation of Sets Via Data Augmentation and Support Vector Machines

Francesca Molinari

(joint work with Haim Bar)

A growing body of literature in econometric theory focuses on estimation and inference in partially identified models such that the identified set of the parameter vector of interest, denoted θ, can be expressed as the zero level set of a non-negative criterion function Q, see [3]:

$$\Theta_I = \{\theta \in \Theta : Q(\theta) = 0\}.$$

In particular, much work has been devoted to develop methodologies that yield confidence sets for the identification region of the model parameters that satisfy various desirable properties, including coverage of each element of the set, see [5], or coverage of the entire set [3] with a prespecified asymptotic probability, possibly uniformly, see [1]. Nonetheless, empirical applications of these methodologies have been hampered by the substantial burden associated with computing the estimated regions and their confidence sets. The only exception is for the case in which the identified set is convex, in which case estimation and inference can be easily carried out using the *support function* of the set, see [2].

This talk aims at extending the empirical applicability of the partial identification approach, including to cases where the parameter vector θ is high dimensional and the identified set is not convex, by providing a simple procedure for computing the regions that exploits learning theory in ways that have not been previously applied in econometrics.

The key insight that leads to our approach is the observation that computing the identification region or its confidence set –with some abuse of notation both denoted Θ_I in what follows– is conceptually a problem of pattern recognition, see [7]. We therefore assume that the set Θ_I is regular closed, and propose to use the following steps:

- Draw θ_i according to a marginal distribution chosen by the user, P_θ on Θ–hence yielding a data augmentation step;
- Obtain its label as

$$y_i = 2 \times 1(Q(\theta_i) = 0) - 1;$$

  This yields an i.i.d. sample ("training data") $D = (\theta_i, y_i)_{i=1}^n$ defined on $(\Theta \times Y)^n$, with $Y := \{-1, 1\}$, and where $P(y|\theta)$ is degenerate because the relation between θ and y is deterministic;
- Obtain a decision function that classifies new instances of θ, i.e., predicts the label y of a new sample (θ, y) drawn from P independently of D, by minimizing an expected loss function (risk);
- Use Support Vector Machines (SVM, see [7]) to obtain the decision function, thereby obtaining a functional form representation of the boundary of Θ_I that can be computed efficiently (via quadratic programming), has good generalization performance, and can be "kernelized" to allow for nonlinear boundaries, see [4].

Using results of [6], we show that the computed identified set converges to the true identified set with respect to the distance in measure, as the training sample size grows to infinity. Under the assumption that the function Q is continuous, we also show that if one slightly modifies the problem by inserting a tolerance δ into the accuracy of the boundary, the convergence rate is of the order of $1/n$. This is achieved by redefining the labels as

$$y_i = 1(Q(\theta_i) = 0) - 1(Q(\theta_i) \geq \delta);$$

and training the SVM only on the data such that $y_i \neq 0$. Intuitively, this modification of the problem puts positive geometric distance between the two classes to be separated and as such makes the task of separation easier.

Preliminary Monte Carlo exercises indicate that the method performs very well in practice, thereby illustrating another application where SVM can be useful in practice.

References

[1] A. Andrews and G. Soares. *Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection* Econometrica **78** (2010), 119–157.

[2] A. Beresteanu, I. Molchanov, and F. Molinari. *Sharp identification regions in models with convex moment predictions* Econometrica **79** (2011), 1785–1821.

[3] V. Chernozhukov, H. Hong, and E. Tamer. *Estimation and Confidence Regions for Parameter Sets in Econometric Models* Econometrica **75** (2007), 1243–1284.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Methods*. Cambridge University Press, 2000.

[5] G.W. Imbens and C.F. Manski. *Confidence Intervals for Partially Identified Parameters* Econometrica **72** (2004), 1845–1857.

[6] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science+Business Media, LLC, 2008.

[7] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1999.

Individual Heterogeneity and Average Welfare

Whitney Newey

(joint work with Jerry A. Hausman)

Demand functions can vary across individuals in general ways. Thus, it is important to allow for general heterogeneity in demand analysis. We consider heterogeneous demand where preferences and linear budget sets are statistically independent. We find that the dimension of heterogeneity and the individual demand functions are not identified. An important purpose of demand analysis is to carry out economic welfare comparisons. Here we find that the exact consumer surplus of a price change, averaged across individuals, is not identified, motivating bounds analysis. We use bounds on income effects to derive relatively simple bounds on the average surplus, including for discrete/continuous choice. We also sketch an approach to bounding surplus that does not use income effect bounds. We apply these results to gasoline demand. We find little sensitivity to the income effect bounds in this application.

Identification and critical dimension in semiparametric estimation

Vladimir Spokoiny
(joint work with Bill E. Xample, Max Muster)

Many statistical tasks can be viewed as problems of semiparametric estimation when the unknown data distribution is described by a high or infinite dimensional parameter while the target is of low dimension. Typical examples are provided by functional estimation, estimation of a function at a point, or simply by estimating a given subvector of the parameter vector. The classical statistical theory provides a general solution to this problem: estimate the full parameter vector by the maximum likelihood method and project the obtained estimate onto the target subspace. This approach is known as *profile maximum likelihood* and it appears to be *semiparametrically efficient* under some mild regularity conditions. We refer to the papers [Murphy and Van der Vaart, 2000, Murphy and Van der Vaart, 1999] and the book [Kosorok, 2005] for a detailed presentation of the modern state of the theory and further references. The famous Wilks result claims that the likelihood ratio test statistic in the semiparametric test problem is nearly chi-square with p degrees of freedom corresponding to the dimension of the target parameter. Various extensions of this result can be found e.g. in [Fan et al., 2001, Fan and Huang, 2005, Boucheron and Massart, 2011]; see also the references therein.

This study revisits the problem of profile semiparametric estimation and addresses some new issues. The most important difference between our approach and the classical theory is a nonasymptotic character of our study. A finite sample analysis is particularly challenging because most of notions, methods and tools in the classical theory are formulated in the asymptotic setup with growing sample size. Only few finite sample general results are available; see e.g. the recent paper [Boucheron and Massart, 2011]. The results of this paper explicitly describes all "small" terms in the expansion of the log-likelihood. This helps to carefully treat the question of applicability of the approach in different situations. A particularly important question is about the critical dimension of the target p and the full parameter dimension p^* for which the main results are still accurate.

We apply the recent bracketing approach of [Spokoiny, 2012] and demonstrate its power on the considered case of semiparametric estimation. Let $\boldsymbol{Y}$ denote the observed random data, and $\boldsymbol{P}$ denote the data distribution. The parametric statistical model assumes that the unknown data distribution $\boldsymbol{P}$ belongs to a given parametric family $(\boldsymbol{P}_{\boldsymbol{\upsilon}})$. The maximum likelihood approach in the parametric estimation suggests to estimate the whole parameter vector $\boldsymbol{\upsilon}$ by maximizing the corresponding log-likelihood $\mathcal{L}(\boldsymbol{\upsilon}) = \log \frac{d\boldsymbol{P}_{\boldsymbol{\upsilon}}}{d\mu_0}(\boldsymbol{Y})$ for some dominating measure μ_0:

$$[c]\tilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\Upsilon} \mathcal{L}(\boldsymbol{\upsilon}).$$

In the semiparametric framework, the target of analysis is only a low dimensional component $\boldsymbol{\theta}$ of the whole parameter $\boldsymbol{\upsilon}$. The *profile maximum likelihood* approach

defines the estimator of $\boldsymbol{\theta}^*$ by projecting the obtained MLE $\tilde{\boldsymbol{\upsilon}}$ on the target space:

$$[c]\tilde{\boldsymbol{\theta}} = P\tilde{\boldsymbol{\upsilon}}.$$

Below we define

$$[c]\breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\substack{\boldsymbol{\upsilon}\in\Upsilon \\ P\boldsymbol{\upsilon}=\boldsymbol{\theta}}} \mathcal{L}(\boldsymbol{\upsilon}).$$

The famous Wilks result can be rewritten as

$$[c]2\bigl\{\breve{L}(\tilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*)\bigr\} \xrightarrow{w} \chi^2_p.$$

The *local asymptotic normality* (LAN) approach by Le Cam leads to the most general setup in which the Wilks type results can be established. The recent paper [Spokoiny, 2012] offers a new look at the classical LAN theory. The basic idea is to replace the local approximation by *local bracketing*. In this paper we show that the local bracketing approach of [Spokoiny, 2012] can be used for obtaining a version of the Wilks Theorem in a quite general semiparametric setup avoiding any special construction like "the hardest parametric submodel"; see [Kosorok, 2005].

Theorem 2. *Let $\boldsymbol{\theta}^*$ be the true target parameter. It holds*

$$[c]\bigl|2\breve{L}(\tilde{\boldsymbol{\theta}}) - 2\breve{L}(\boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\bigr| \le \mathtt{C}\,\tau_{\boldsymbol{\epsilon}} p^*,$$

where p^* is the full parameter dimension, $\tau_{\boldsymbol{\epsilon}}$ is a small constant, and $\breve{\boldsymbol{\xi}}$ is a random p-vector satisfying $\mathbb{E}\breve{\boldsymbol{\xi}} = 0$ and $\mathbb{E}\|\breve{\boldsymbol{\xi}}\|^2 \cong p$. Moreover, deviation properties of $\|\breve{\boldsymbol{\xi}}\|^2$ resemble the ones of a chi-square random variable with p degrees of freedom.

In the i.i.d. case this implies

Theorem 3. *Let $Y_1, \ldots, Y_n$ be i.i.d. $\boldsymbol{P}_{\boldsymbol{\upsilon}^*}$. If*

$$[c]\beta_n \stackrel{\text{def}}{=} p^{*3/2}/n^{1/2},$$

then it holds with a dominating probability:

$$\begin{aligned}
[ccl]\bigl\|(n\breve{\mathbb{F}})^{1/2}\bigl(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}\bigr\|^2 &\le \mathtt{C}\beta_n, \\
\bigl|2\breve{L}(\tilde{\boldsymbol{\theta}}) - 2\breve{L}(\boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2\bigr| &\le \mathtt{C}\beta_n.
\end{aligned}$$

Moreover, the p-vector $\breve{\boldsymbol{\xi}} \stackrel{\text{def}}{=} \breve{\mathbb{F}}^{-1/2}\bigl(\nabla_{\boldsymbol{\theta}} - \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}\mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\nabla_{\boldsymbol{\eta}}\bigr)$ is asymptotically standard normal as $n \to \infty$. Here $\mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}$ and $\mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}$ are blocks of the Fisher information matrix, while $\breve{\mathbb{F}}$ is the relative Fisher information matrix for $\boldsymbol{\theta}$. This yields the asymptotic efficiency of the profile MLE $\tilde{\boldsymbol{\theta}}$.

A special example for a Poisson model shows that the condition $\beta_n^2 = p^{*3}/n \to 0$ is necessary for the Wilks result and cannot be relaxed or dropped.

References

[Boucheron and Massart, 2011] Boucheron, S. and Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150:405–433. 10.1007/s00440-010-0278-7.

[Fan and Huang, 2005] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057.

[Fan et al., 2001] Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, 29(1):153–193.

[Kosorok, 2005] Kosorok, M. (2005). *Introduction to Empirical Processes and Semiparametric Inference*. Springer in Statistics.

[Murphy and Van der Vaart, 1999] Murphy, S. A. and Van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli*, 5(3):381–412.

[Murphy and Van der Vaart, 2000] Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

[Spokoiny, 2012] Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.

Principal Component Analysis in an Asymmetric Norm

Ngoc Mai Tran

(joint work with Maria Osipenko, Wolfgang Härdle)

Principal component analysis (PCA) is a widely used dimension reduction tool in the analysis of many kind of high-dimensional data. However, in many of the above applications, one is interested in capturing the tail of the data rather than the mean. In this paper, we develop an analgoue of PCA for quantiles and expectiles. The difficulty is that there is no natural basis, no 'principal components', to the k-dimensional subspace found. We propose two definitions of principal components, provide algorithms based on iterative least squares. We prove upperbounds on their convergence times, and compare their performances in practice using the Chinese Weather dataset.

When data come as curves without known functional form, the statistician faces immediately the need for dimension reduction. The conventional and widely used tool for such high dimensional curve data is principal component analysis (PCA). The basic principle of this technique is to treat the curves as random variations around a mean curve, and then orthogonalize the covariance operator into eigenfunctions and corresponding (random) loadings. The focus of such a representation is on studying the variation around a mean curve. Loadings on (interpretable) eigenfunctions would then represent specific variations around the average. PCA or more generally functional PCA (FPCA) has been successfully applied in many fields such as gene expression measurements, financial with factor analysis, weather and natural hazard studies, demographics, etc... One of the first applications is the one reported in [5]. They considered temperature curves recorded daily over a year at multiple stations in an area. The premise is that there are only a few principal components influencing the average temperature, and that the temperature curve from each station is well-approximated on average by a specific linear combinations of these factors. PCA approximates the mean of

the data by a nested sequence of optimal subspaces of small dimensions. Thus the optimal subspace of dimension k comes with a natural basis, consisting of uncorrelated random curves (vectors), the principal components, playing the role of the factors aforementioned. Due to the nested structure of the optimal subspaces, one can compute the first few components using a greedy algorithm. The first principal component can be computed efficiently using iterative partial least squares [8].

In many of the above applications, one is not only interested in the variation around an average curve, but rather in features of the data that are expressible as scale (variance) or tail related functional data. In volatility related pricing of financial products, for example, the variation of the scale of risk factors is at the core of fair pricing. If one would like to construct weather derivatives or forecasts for the above FPCA example on temperature curves, one needs not only to know the variation across stations, but also the changing scale of the temperature curves, [1], [6]. In climatological science, one is interested in the extremes of certain natural phenomena like drought or rainfall. A tail indicator like a quantile of a conditional distribution when indexed by an explanatory variable also constitutes a curve. Therefore, such a quantile curve collection may also be treated in a FPCA context. Yet another tail-describing curve is the expectile function. Like the quantile curve, it can be represented via a solution with respect to an asymmetric norm.

In this paper, we develop an analogue of PCA for quantiles and expectiles. The later, proposed by [4], is an analogue of the mean for quantiles. The quantile to level τ of a distribution with cdf F, assuming F is invertible, is defined as $q_\tau = F^{-1}(\tau)$. It is also the solution to the following optimization problem [4]

$$q_\tau = \arg\min_{q\in\mathbb{R}^p} \mathbb{E}\|X - q\|_{\tau,1}$$

where X is a random variable with distribution F, and

$$\|x\|_{\tau,1} = |\mathbf{1}(x \le 0) - \tau||x|^\alpha, \quad \alpha = 1. \tag{1}$$

Given data $X_i \sim F, i = 1, \ldots, n$, one may formulate the estimation of the unknown quantile in a location model:

$$X_i = q_\tau + \epsilon_i, \tag{2}$$

with τ-quantile of the CDF of ϵ being zero. A natural estimate of q_τ in (2) is therefore

$$\hat{q}_\tau = \arg\min_{q\in\mathbb{R}^p} \sum_{i=1}^{n} \|X_i - q\|_{\tau,1}. \tag{3}$$

Formulation (3) yields a statistical interpretation. In fact, if the noise ϵ_i in (2) follow a so-called asymmetric Laplace distribution $ALD(\tau)$, which has cdf proportional to $\exp(-\rho_\tau(\cdot))$, then (3) can be interpreted as a quasi likelihood estimation equation of (2). Putting $\alpha = 2$ in 1 yields, via (3), a quasi likelihood interpretation based on an asymmetric normal distribution.

As noted in [3], the first step in this problem corresponds to doing low-rank matrix approximation with weighted ℓ_1 and ℓ_2 norm, respectively, where the weights are sign-sensitive (see Section 1). Based on a proposal of [7], the authors of [3] proposed an iterative weighted least squares algorithm for expectiles, where the weights are updated in each iteration. This algorithm is guaranteed to converge, although not necessarily to the global minimum as we shall show below. Thus one can at least find a locally optimal k-dimensional subspace that best approximates a given quantile or expectile. The difficulty is that the weight matrix is not of rank one, hence there is no natural basis, no 'principal components', to the k-dimensional subspace found. While this is a known problem in weighted low-rank matrix approximation [8], this problem has not been addressed in [3]. Furthermore, the definition of optimal τ-expectile subspace employed in [3] is not invariant under linear transformations of the data. That is, if one changes the basis of the data, the optimal τ-expectile subspace in the new basis is not necessarily a linear transform of that expressed in the old basis. This means one has to fix a basis for the data before computing the optimal τ-expectile subspace. This restricts the usefulness of this method to applications where there is a natural basis, such as in the Chinese weather dataset, where yearly temperature is expressed as a vector of 365 daily temperatures. Here one would be interested in capturing extreme daily temperature as opposed to extreme temperature expressed in a Fourier basis. However, in many other applications, invariance under change of basis is an important feature of PCA.

The contributions of our paper is two fold. First, we work with the formulation in [3] and propose two natural bases, hence two definitions of principal components for the optimal subspace found. Second, we propose an alternative definition of principal components for quantiles and expectiles, closely related to the definition of principal directions for quantiles of [2]. This definition satisfies many nice properties, such as invariance under translations and linear transformations of the data, and in particular, returns the usual PCA basis under elliptically symmetric distributions. We then provide algorithms to compute the three versions principal components aforementioned, based on iterative weighted least squares. We prove upper bounds on their convergence times, and compare their performances in practice using the Chinese weather dataset.

References

[1] S. D. Campbell and F. X. Diebold, *Weather forecasting for weather derivatives*, Journal of the American Statistical Association **100(469)** (2005), 6-16.

[2] R. Fraiman and B. Pateiro-Lòpez, *Quantiles for finite and infinite dimensional data*, Journal of Multivariate Analysis **108** (2012), 1-14.

[3] M. Guo, L. Zhou, W. K. Härdle and J. Huang *Functional data analysis for generalized quantile regression*, in Sonderforschunfsbereich 649: Ökonomisches Risiko-(SFB 649 Papers), Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät (2012).

[4] W. K. Newey and J. P. Powell, *Asymmetric least squares estimation and testing*, Econometrica: Journal of the Econometric Society **55(4)** (1987), 819-847.

[5] J. O. Ramsay and B. W. Silverman, *Functional data analysis* (2005), Springer, New York.

[6] J. S. Benth and F. E. Benth, *A critical view on temperature modelling for application in weather derivatives markets*, Energy Economics **34(2)** (2012), 592-602.
[7] S. K. Schnabel, *Expectile smoothing: new perspectives on asymmetric least squares: an application to life expectancy*, Sociale Wetenschappen Proefschriften, Dissertation (2011).
[8] N. Srebro and T. Jaakkola, *Weighted low-rank approximations*, in Machine Learning International Workshop **20(2)** (2003), 720-728.

Composite Quantile Regression for the Single-Index Model

Weining Wang
(joint work with Yan Fan, Wolfgang Karl Härdle, and Lixing Zhu)

Regression between response Y and covariates X is a standard element of statistical data analysis. When the regression function is supposed to be estimated in a nonparametric context, the dimensionality of X plays a crucial role. Among the many dimension reduction techniques the single index approach has a unique feature: the index that yields interpretability and low dimension simultaneously. In the case of ultra high dimensional regressors X though it suffers, as any regression method, from singularity issues. Efficient variable selection is here the strategy to employ. Specifically we consider a composite regression with general weighted loss and possibly ultra high dimensional variables. Our setup is general, and includes quantile, expectile (and therefore mean) regression. We offer theoretical properties and demonstrate our method with applications to firm risk analysis in a CoVaR context.

Quantile regression(QR) is one of the major statistical tools and is "gradually developing into a comprehensive strategy for completing the regression prediction" [13]. In many fields of applications like quantitative finance, econometrics, marketing and also in medical and biological sciences, QR is a fundamental element for data analysis, modeling and inference. An application in finance is the analysis of conditional Value-at-Risk (VaR). [5] proposed the CaViaR framework to model VaR dynamically. [12] used their QR techniques to test heteroscedasticity in the field of labor market discrimination. Like expectile analysis it models the conditional tail behavior.

The QR estimation implicitly assumes an asymmetric ALD (asymmetric Laplace distribution) likelihood, and may not be efficient in the QMLE case. Therefore, different types of flexible loss functions are considered in the literature to improve the estimation efficiency, such as, composite quantile regression, [29], [9] and [10]. Moreover, [3] proposed a general loss function framework for linear models, with a weighted sum of different kinds of loss functions, and the weights are selected to be data driven. Another special type of loss considered in [17] corresponds to expectile regression (ER) that is in spirit similar to QR but contains mean regression as its special case. Nonparametric expectile smoothing work with application to demography could be found in [19]. The ER curves are alternatives to the QR curves and give us an alternative picture of regression of Y on X.

The difficulty of characterizing an entire distribution partly arises from the high dimensionality of covariates, which asks for striking a balance between model flexibility and statistical precision. To crack this tough nut, dimension reduction techniques of semiparametric type such as the single index model came into the focus of statistical modeling. [23] considered quantile regression via a single index model. However, to our knowledge there are no further literatures on generalized QR for the single-index model.

In addition to the dimension reduction, there is however the problem of choosing the right variables for projection. This motivates our second goal of this research: variable selection. [14], [22] and [27] focused on variable selection in mean regression for the single index model. Considering the uncertainty on the multi-index model structure, we restrict ourselves to the single-index model at the moment. An application of our research is presented in the relevant financial risk area: to investigate how the revenue distribution of companies depends on financial ratios describing risk factors for possible failure. Such kind of research has important consequences for rating and credit scoring.

When the dimension of X is high, severe nonlinear dependencies between X and the expectile (quantile) curves are expected. This triggers the nonparametric approach, but in its full gear, it runs into the "curse of dimensionality" trap, meaning that the convergence rate of the smoothing techniques is so slow that it is actually impractical to use in such situations. A balanced dimension reduction space for quantile regression is therefore needed. The MAVE technique, [24] provides us 1) with a dimension reduction and 2) good numerical properties for semiparametric function estimation. The set of ideas presented there, however, have never been applied to composite quantile framework or an even more general composite quasi-likelihood framework. The semiparametric multi-index approach that we consider herein will provide practitioners with a tool that combines flexibility in modeling with applicability for even very high dimensional data. Consequently the curse of dimensionality is circumvented. The Lasso idea in combination with the minimum average contrast estimate (MACE) technique will provide a set of relevant practical techniques for a wide range of disciplines. The algorithms used in this project are published on the quantlet database www.quantlet.org.

References

[1] Adrian, T. and Brunnermeier, M. K. (2011). CoVaR. *Staff Reports 348, Federal Reserve Bank of New York*.

[2] Berkowitz, J., Christoffersen, P. and Pelletier, D. (2009). Evaluating value-at-risk models with desk-level data. *Working Paper 010, North Carolina State University, Department of Economics*.

[3] Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B.* **73** (3) 325–349.

[4] Chao, S. K., Härdle, W. K. and Wang, W. (2012). *Quantile regression in Risk Calibration. In Handbook for Financial Econometrics and Statistics (Cheng-Few Lee, ed.)*. Springer Verlag, forthcoming, SFB 649 DP 2012-006.

[5] Engle, R. F. and Manganelli, S. (2004). CaViaR: Conditional autoregressive value at risk by regression quantiles. *J. Bus. Econ. Stat.*. **22** 367–381.

[6] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.*. **96** 1348–1360.
[7] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*. **84** 986–995.
[8] Huber, P. J. (1985). Projection pursuit. *Ann. Math. Statist.*. **13** 435–475.
[9] Kai, B., Li, R. and Zou, H. (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *J. R. Statist. Soc. B*. **72** 49–69.
[10] Kai, B., Li, R. and Zou, H. (2011) New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *Ann. Statist.*. **39** (1) 305–332.
[11] Koenker, R. and Bassett, G. W. (1978). Regression quantiles. *Econometrica*. **46** 33–50.
[12] Koenker, R. and Bassett, G. W. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*. **50** 43–61.
[13] Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Econometric Perspectives*. **15** (4) 143–156.
[14] Kong, E. and Xia, Y. (1994). Variable selection for the single-index model. *Biometrika*. **94** 217–229.
[15] Leng, C., Xia, Y. and Xu, J. (2008). An adaptive estimation method for semiparametric models and dimension reduction. *WSPC-Proceedings*.
[16] Li, Y. and Zhu, J. (2008). L1- norm quantile regression. *J. Comput. Graph. Stat.*. **17** 163–185.
[17] Newey, W. and Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*. **55** 819–847.
[18] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*. **90** 1257–1270.
[19] Schnabel, S. and Eilers, P. (2009). Optimal expectile smoothing. *Comput. Stat. Data. An.*. **53** (12) 4168–4177.
[20] Serfling, R. J. (2001). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
[21] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*. **58** (1) 267–288.
[22] Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection methods for high-dimensional data: sparse mave. *Comput. Stat. Data. An.*. **52** 4512–4520.
[23] Wu, T. Z., Yu, K. and Yu, Y. (2010). Single-index quantile regression. *J. Multivariate Anal.*. **101** 1607–1621.
[24] Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Statist. Soc. B*. **64** 363–410.
[25] Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.*. **93** 228–237.
[26] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*. **68** (1) 49–67.
[27] Zeng, P., He, T. H. and Zhu, Y. (2012). A lasso-type approach for estimation and variable selection in single-index models. *J. Comput. Graph. Stat.*. **21** 92–109.
[28] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*. **101** 476.
[29] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.*. **36** (3) 1108–1126.

Reporter: Maria Grith

Participants

Prof. Dr. Federico A. Bugni
Department of Economics
Duke University
Social Sciences Building
Box 90097
Durham NC 27708-0097
UNITED STATES

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
8092 Zürich
SWITZERLAND

Prof. Dr. Ivan Canay
Northwestern University
Department of Economics
2001 Sheridan Road
Evanston, IL 60208-2600
UNITED STATES

Prof. Dr. Andrew Chesher
Department of Economics
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Prof. Dr. Alfred Galichon
Departement de Economique
Sciences Po
28 Rue des Saint-Peres
75007 Paris
FRANCE

Maria Grith
Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität zu Berlin
Spandauer Str. 1
10178 Berlin
GERMANY

Prof. Dr. Wolfgang Karl Härdle
Wirtschaftswissenschaftliche Fakultät
Ladislaus v. Bortkiewicz Chair of
Statistics CASE
Humboldt-Universität zu Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Prof. Dr. Marc Henry
Departement de Sciences Economiques
Universite de Montreal
Pavillon Lionel-Groulx
3150, rue Jean-Brillant
Montreal Quebec H3T 1N8
CANADA

Prof. Dr. Joel L. Horowitz
Northwestern University
Department of Economics
2001 Sheridan Road
Evanston, IL 60208-2600
UNITED STATES

Prof. Dr. Tatyana Krivobokova
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen
GERMANY

Prof. Dr. Sokbae Lee
Department of Economics
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul 151-742
KOREA, REPUBLIC OF

Prof. Dr. Yuan Liao
Department of Mathematics
University of Maryland
College Park, MD 20742-4015
UNITED STATES

Prof. Dr. Enno Mammen
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68161 Mannheim
GERMANY

Prof. Dr. Nicolai Meinshausen
Department of Statistics
Oxford University
1 South Parks Road
Oxford OX1 3TG
UNITED KINGDOM

Prof. Dr. Ilya S. Molchanov
Institut f. Mathematische Statistik
Universität Bern
Sidlerstr. 5
3012 Bern
SWITZERLAND

Prof. Dr. Francesca Molinari
Department of Economics
Cornell University
402 Uris Hall
Ithaca, NY 14853-7601
UNITED STATES

Prof. Dr. Whitney K. Newey
Department of Economics
Massachusetts Institute of
Technology
50 Memorial Drive
Cambridge, MA 02142-1347
UNITED STATES

Prof. Dr. Yaacov Ritov
Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905
ISRAEL

Prof. Dr. Adam M. Rosen
Department of Economics
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Vladimir G. Spokoiny
Weierstrass-Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
10117 Berlin
GERMANY

Dr. Ngoc Mai Tran
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

Weining Wang
Humboldt-Universität zu Berlin
Ladislaus v. Bortkiewicz Chair of Stat.
School of Business and Economics
Unter den Linden 6
10099 Berlin
GERMANY