A Primer on Structural Estimation in Accounting Research*

JEREMY BERTOMEU YING LIANG IVÁN MARINOVIC

Abstract

This primer offers a hands-on accessible guide to writing and estimating structural models. We review commonly-used methodologies, including dynamic programming, maximum likelihood, generalized and simulated method of moments, conditional choice probabilities as well as tools to compute standard errors and common diagnostics and tests of economic hypotheses. Special attention is devoted to the bootstrap as a convenient toolbox to estimate complex economic interactions. The methods are illustrated with recent developments in earnings management, auditing, investment, conservatism, and financial disclosures. Intuition and applications are emphasized over formalism.

*Jeremy Bertomeu is an Associate Professor at Olin Business School, Washington University, Saint Louis. Address: E. Cheynel, Olin Business School, Washington University, Snow Way, 1 Brookings Drive, Saint. Louis, MO 63130. Contact author: bjeremy@wustl.edu. Ying Liang is an Assistant Professor at J. Mack Robinson College of Business, Georgia State University. Address: 35 Broad St NW, Atlanta, GA 30303. Iván Marinovic is an Associate Professor at Stanford Graduate School of Business. Address: 655 Knight Way, Stanford, CA 94305. A github repository is maintained with all the estimation code used in this survey at https://github.com/yingliang888/survey.

Structural estimation refers to a class of empirical models in which the economic assumptions describing decision problems are formally stated and used to mathematically derive the empirical model. The method has seen unprecedented growth in accounting, having the potential to answer questions of causal inference that are difficult to answer with traditional reduced-form modelling (Bertomeu, Beyer and Taylor 2015, Gow, Larcker and Reiss 2016). Continuing progress in developing theories more amenable to describe real data and a trend toward more efficient computing, but also the wide body of knowledge established from its use in other areas of the social sciences including economics, finance and marketing, provide opportunities to draw new empirical and theoretical insight from the method.

While this monograph will be primarily focused on developing practical tools, it is nevertheless useful to reflect on the fundamental premises that underlie what the method intends to achieve. As in the natural sciences, structural modelling postulates the existence of stable laws governing economic reality. These laws are expressed mathematically (i.e., a set of equations) as a function of both observable and unobservable factors. However, unlike, for instance, the law of universal gravitation, economic laws are highly complex, potentially involving a large number of unobservable variables. A full apprehension of the underlying law is impossible and even useless, for the same reason a full scale map would be useless (Borges 1998). Hence, the structural approach formulates stylized (miniature) representations of the underlying laws which –despite their stylized nature— are expected to shed light on a specific, but hopefully important, aspect of economic reality. Assumptions are needed to isolate and characterize the causal impact of a single factor.

As Cartwright (2007) notes, the stylized nature of the theory, understood as a deviation from descriptive realism, is part of the research paradigm. Indeed, the high degree of idealization is essential to the ability of the model to reveal the real world.[1] To begin, the

[1]The problem is the presence of "false assumptions" that provide internal validity but distort our deductions, something that, according to Cartright, pervades economic modelling. She argues that in economics we achieve empirical identification via assumptions that are not widely accepted, such as the principles of

researcher formulates a theory model, consisting of a set of assumptions. As Marschak (1974) puts it, a model is "1) a set of relations describing human behavior and institutions, as well as technological laws, and generally involving non-observable random disturbances and non-observable random errors in measurement, 2) the joint probability distribution of these random quantities." These assumptions are sufficient to generate a set of empirical predictions which, when the model is identified, are sensitive to the model's parameters. The model's predictions are thus contrasted with their empirical analogues, via statistical analysis, to produce parameter estimates.

For some questions, this approach offers several advantages relative to non-structural approaches. In his memorial Lectures, Koopmans (1975) argues that "the decision not to use theory of man's economic behavior limits the value to economic science and to the maker of policies, of the results obtained [by empirical methods]." First, the model imposes restrictions on the data that satisfy sound economic principles, requiring the researcher to be explicit about what is assumed or ruled out. These assumptions discipline the analysis enforcing coherence throughout the study, which is hard to implement in the absence of formal theory, as this often leads to a situation where each hypothesis relies on a different set of potentially inconsistent assumptions.

Second, because the model aims to identify economic primitives that are exogenous to the mechanisms of the model, the researcher can answer "what if" questions and evaluate the consequences of such changes to the environment (also known as counter-factuals). The objective is to estimate primitives, defined as parameters that are invariant to changes in other parameters or policies within the scope of a research problem, and serves to guide policy makers interested in understanding the effect of policy changes (e.g., the effect of minimum salary on unemployment).

utility theory, but *ad hoc* and controversial. Structural models are also an opportunity to address this criticism by making explicit what the assumptions are, and, for suitable datasets, allow for departures guided by empirical evidence, i.e., violations of Bayesian updating, behavioral biases, or preferences other than standard expected utility.

Third, there is no other approach that can offer a complete empirical description of an economic model: as a scientific exercise, this approach is an effort to quantitatively uncover (an aspect of) laws organizing the data. Since all theoretical implications are spelled out, this approach often allows us to reject theories that are not consistent with the data, even if on the surface they might seem so. Thus, the structural approach satisfies the riskiness criterion that Popper (2014) highlights as the defining feature of a genuine science: a structural model typically offers many experiments that can (and often) lead to the rejection of the underlying theory. This is important: one of the benefits of the structural approach is that, as a researcher, one is never left empty-handed: rejecting a meaningful theory is itself an important discovery that requires creativity and often leads to new and more realistic theories.

The problem of using data to quantify the predictions of economic models is, of course, not new. Stepping outside of social sciences, almost all models in the hard sciences are structural. Models of epidemiology, for example, are based on assumptions about the spread of infectious diseases; in physics, models of particles satisfy primitives about the standard model. Likewise, in macroeconomics, authors have used structural models to pin down the effect of monetary and fiscal policy long before it was realistic or computationally feasible to use statistical methods (Kydland and Prescott 1982). Rather than a tool, structural models are a philosophical viewpoint using theory to organize data starting from assumptions about laws of nature.

Having noted these broader objectives, the monograph aims to provide an introductory primer to researchers interested in incorporating structural models into their analysis. The essay is designed for researchers with little or no prior knowledge of structural models, and with the objective to make technical barriers to entry into this literature no greater than in other theoretical or empirical exercises. The emphasis is on adequate use of the methods in applied work. Readers interested in these issues will find many textbook references in text with various developments, formal analyses and proofs. While most examples are

drawn from accounting research, many of the methods can be (and have been) applied more generally to other related areas such as finance, marketing, and economics.

A theme developed throughout is that substance is more important than form. By substance, one means bringing into the model economic mechanisms or questions that were, prior, not fully resolved by theory or reduced-form empirical work. Tools can help answer a richer question but are always means to an end. It is rare to find a "better" tool that does not have its own limitation. A larger model, i.e., which includes more economic trade-offs, can be more obscure, require more technical assumptions, be less computationally stable or hard to replicate, and require more data. An asymptotically more accurate estimator may be more sensitive to misspecification and unnecessarily complicated when simpler approaches are sufficiently precise for the question of interest. In summary, a better model is not a more general model, but one that gives a persuasive robust answer to the question after considering *all* technical and empirical challenges. To this effect, the objective is to develop a set of methods that can be applied over a variety of contexts, so that one may choose the simplest more transparent tool for the question.

Whited (2021) offers the following common objectives for a structural model, ranked from harder to easier for typical applications. The first objective is to estimate the parameters of interest to understand an empirical setting. The second objective is to falsify a theory in order to help find a better theory. The third objective is to run counterfactual "what if" exercises if a particular course of actions were undertaken. To these important functions, we add a fourth broader objective that subsumes all three: to provide scientists with a plausible and internally-consistent representation of economic reality that unifies theory and data.

The first objective is probably the most difficult. Most structural models in social sciences do not aim to offer a descriptively accurate representations of an empirical setting. They are simplifications to reduce a complex set of interactions to important first-order considerations. The interpretation of estimated parameters is usually not as primitive

laws of nature but in terms of their implications about decision-making. Nevertheless, for classic models of broad general interest, certain parameters may be economically meaningful even in a simplified model. For example, the estimated bankruptcy cost and cost of external finance in Hennessy and Whited (2007) are parameters fundamental in Modigliani-Miller theorems, learning about managerial ability is fundamental in theories of endogenous turnover (Taylor 2010), the frictions to unravelling in Bertomeu, Ma and Marinovic (2020) are foundational in disclosure theory, and the cost of shirking estimated in Gayle and Miller (2005) underlie all agency theory.

The second objective usually takes the form of assessing the fit of a model or comparing a model against another. It is a scientific process to identify areas of disagreement between model and data in order to guide model choice and potential improvements. In asset pricing, the equity premium puzzle of Mehra and Prescott (1985) led to considerable innovation in model building incorporating richer preferences and types of risks. But assessing theory using a structural model is subject to caution because test statistics used for model diagnostics are joint tests of economic assumptions and many ancillary technical assumptions, such as functional forms and unobserved heterogeneity.

The third objective is to offer a quantitative assessment of the economic consequences of a policy decision *that has not yet been made* and, hence, for which data does not exist. A regulator may wish to assess the effect of a change in regulation which, with reduced-form analysis only, would require to conduct randomized trials. Taking aside the potential costs and fairness of such experimentation, for many firm-level questions in accounting, randomized trials are infeasible because all firms are inter-connected and an experiment on one set of firms would affect other firms held as controls. Structural models do not provide the same level of certainty as an *ideal* randomized trial, but they are the only tool available to conduct such policy "what if" experiments when experimentation are unavailable. Indeed, random assignments can be less effective at providing useful counter-factuals than a structural model when the random assignment destroys choices

made empirically (for example, signalling or information acquisition) and no longer represent the same environment (Hennessy and Chemla 2022)

The process of writing a structural model is unusual compared to other methods because it involves a back and forth dialogue between theory and data. This process can be daunting and, without organization, may involve restarting a project multiple times and wasting valuable insights because of unclear diagnostics on the parts of the model that work versus those that fail. Hence, some researchers may find it useful to organize a workflow to decompose the analysis into smaller steps and features decision points that require revision to parts of the model or the method used to resolve it.

Step 1: Know the available data. Unlike traditional theory, not all questions will be answerable with a structural model because identification, a problem that we discuss in more detail later on, is a function of the information contained in the data. The structural model will draw connections between observable and unobservable empirical elements so, ideally, the objective is to identify which data elements should be in the model, what the research question is about and, critically, whether the data will be likely sufficient to answer this question.

To know the data, a useful preliminary step is to approach the question as one to be answered in reduced-form. At this preliminary stage, the researcher is using qualitative models but placing structure only on observables and noise terms, not on the original decision problem. To what extent can the question be resolved from features of the data? Are there stylized facts suggesting that the theory is plausible? What theories are thoroughly incompatible with these stylized facts? The objective of this critical first step is to limit the question to structural models suitable to the empirical sample: these problems are usually at the intersection of stylized facts suggesting that the theory is adequate but with open questions unanswered by the reduced-form analysis.

Step 2: Write a preliminary theoretical model, usually (but not always) by simplifying a theoretical bookshelf model from the relevant theoretical literature. This model need not all have all the checks and balances of a theoretical model because, unlike formal theory which aims to be used conceptually across applications, its analysis will be facilitated by institutional details in a sample. However, the model should capture the important empirical observables and contain a plausible first-order effect.

This step is often the most misunderstood in structural models. Like any scientific exercise, the modeler aims to offer an improvement over our current understanding but is not considering a realistic or complete description of all forces. Therefore, the model is not an attempt at realism; in fact, orthogonal error terms in the statistical model will serve to capture factors that were omitted from the analysis. The preliminary model serves to organize these thoughts into a set of coherent forces and provides the researcher with a conceptual understanding of how the model organizes the data. As an example of this approach, Bertomeu et al. (2020) estimate the static disclosure model in Dye (1985) and Jung and Kwon (1988) before overlaying a full multi-period model with endogenous learning about disclosure frictions.

Step 3: Solve the model either numerically or analytically, and check the main quantitative properties of this model against related features of the data. Usually, the requirement for a model to be amenable for structural is that its variables should not be too stylized and that quantities correspond to those observed. For example, a model with high-level implications about the social value of information and with discrete outcomes may not be well-suited to continuous data in a narrow institutional setting. To know if a model is amenable for structural analysis, a good question to ask is: how many ad-hoc additional assumptions or data interpretations are required to map theoretical constructs into empirical ones?

When using full-solution methods, the model should be solvable in reasonable time,

robustly over many parameters without manual adjustments. The time required to solve the model may also guide the appropriate estimation method: a model that is quick to solve can always be estimated by minimizing a distance between simulated model features and data features. In a typical estimation, the model will be evaluated over many parameter values, so a robust solving algorithm is often required.

Step 4: Revise the model to capture first-order effects that may not be theoretically interesting, in that they may not bring new intuition, but whose interactions with the research problem may substantively affect the estimation. Naturally, the problem here is not to model all possible realistic forces (hence, the important word "substantively") but to think about components with essential interactions.

Is the problem irreducibly dynamic or can the first-order effects be seen in a static model, possibly using different data subsets? Are there observable firm characteristics endogenous to the forces of the model? Is there unobserved heterogeneity that should be written down in the model? A bigger model is a not a better model, because incorporating more elements also involves making additional assumptions and may require more data. Hence, this step should focus on the elements that are feasible and essential to use the model in an empirical setting.

Step 5: Ensure that the revised model can be solved, or has useable restrictions. There exists two approaches to estimating structural models that we will discuss in greater detail in text: (5a) solve the model completely (full solution methods) and derive either the data likelihood or moments according to the model - in the latter case, if the model restrictions are not in closed-form, these model moments can be obtained by simulating data from the model; (5b) write theoretical restrictions from the model, which usually are constraints on the decision problems individuals solve and implications from optimality conditions.

The full solution method is the conceptually simpler approach given that, if the model

can be solved numerically, an estimation procedure can be obtained by computing (numerically) economic features of interest in the model, and find parameters that best match these model predictions to data. Full solution methods combined with selecting adequate features can also achieve more precise estimates, because they use all the optimality implications of the model. They are also theoretically more transparent in principle because the researcher can assess the behavior of the model by simulation; for example, Zakolyukina (2018) estimates a dynamic model of earnings management and, before the estimation, plots manipulation choices predicted by the model as a function of the book value. This preliminary analysis prior to estimation can open the black box of a complex theoretical model. On the other hand, full solution method can be computationally-intensive and, as a result, the implementation of these methods is usually for parsimonious models with few parameters. Recent developments in computing have nevertheless dramatically expanded the scope of problems solvable with full solution methods.

A different approach is to use theoretical restrictions from the model, which may be a subset of the theoretical restrictions. Many endogenous objects in the model, which could be (in principle) solved as a function of model parameters, are observable empirically as individuals or firms making optimal choices. So, rather than solving the model, one can substitute in empirical estimates of endogenous objects and then identify parameters of interest from theoretical restrictions on these objects. The consumption Euler moment condition in asset pricing (Hansen and Singleton 1982, Rust 1994), which link current and future consumption, is a classic example of this method. Conditional choice probabilities are another example in which the value function can be written in terms of observables. Gerakos and Syverson (2015) and Cheynel and Zhou (2020a) are recent applications that estimate client preferences for auditors from observed auditor replacements. These methods are usually computationally more accessible and thus can allow for more richness in models.

Step 6: Fit the model, that is, program code that finds parameter values that ensure that the theoretical restrictions from steps (5a) and (5b) are best met by the data. In the context of a full solution method, the usual approach is to simulate data from the model with various parameter values until the dimensions of interest least distinguish data and model; put differently, knowing how to solve a model numerically is in principle sufficient to be able to fit the model. To be checked is whether the fit economically explains the data: is the model consistent with the motivating facts? Note that a parsimonious model may not always pass a statistical test assessing whether the model is a complete explanation of the data, but should nevertheless generate magnitudes consistent with the data on the main constructs of interest.

There are many diagnostics informative about the performance of a model but one should note that rewriting a model until it passes a test implies that the asymptotic test statistics are no longer valid. Textbook asymptotic distributions of test statistics do not hold if the researcher systematically searches for a model to pass a test, and therefore the statistical meaning of a test statistic p-value is lost. Put differently, the researcher should enter a diagnostics step with a plausible model, and, given that all models will feature some degree of hypothesis testing, the diagnostic should be read in terms of a performance score rather than a binary pass-or-fail.

Step 7: And, of course, answer the research question. This can be a measurement of a hidden quantity of interest and, often, a counter-factual analysis: how would quantities relevant to firms and individuals change in response to different parameters or a new policy? This usually involves changing one part of the model while keeping all the remaining parameter estimates as given, and solving the model numerically with this change. A counter-factual can be a policy that changes the rules of the game, a change in a parameter, or an application of the estimates to a setting with less data.

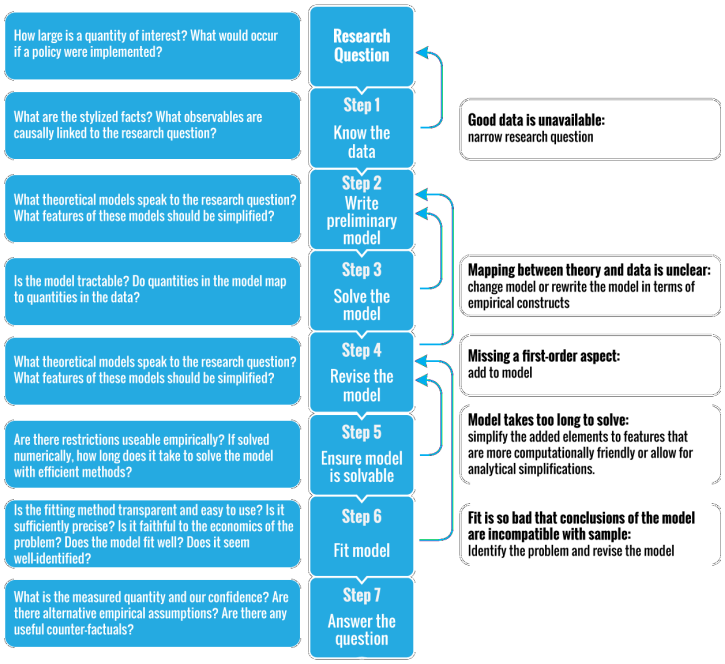


Figure 1: Structural Models: A Workflow

To illustrate all of these steps in a single application, consider the model by Beyer, Guttman and Marinovic (2019), which aims to estimate the noise in reported earnings caused by earnings management. The data used is prices and earnings, over a panel of firms (step 1). There are two bookshelf models that speak to a relation between noisy earnings management and price responses (step 2) in Fischer and Verrecchia (2000) and Dye and Sridhar (2004). Simplifying the Dye and Sridhar (2004) model to a single costly reporting yields an equation in which the earnings response to earnings is a function of the fundamental uncertainty and the earnings management driven uncertainty (step 3). However, mapping this model to data is problematic because the model assumes that agents report the value of the firm, while (in practice) firms report periodic earnings. Answering the question requires to be explicit about value as a dynamic sequence of reported earnings.

Beyer et al. (2019) rewrite the model as a repeated sequence of manipulation choices, adapting the static model (step 4). Fortunately, as is common in linear updating models,

the same guess-and-verify methods to solve a single-period model can be applied to a dynamic model and therefore, the model implies a relation between prices as well as current earnings and lagged earnings and prices (step 5). This relation can then be estimated to recover the economic primitives (step 6). The last step is to measure the amount of uncertainty due to earnings noise or, equivalently, how much pricing error would be removed in a counter-factual where enforcement against manipulation is perfect, and yields that the noise due to earnings management is about half of the fundamental uncertainty (step 7).

This primer is divided into eight sections, which are inter-connected but can be also read in isolation. Section 1 presents two simple guided examples of structural estimation exercises, in which the logic of the main tools can be absorbed with minimal formalism. Section 2 presents a step-by-step approach to structural estimation, generalizing the methods applied in the two examples. Section 3 discusses more details of the econometric methods for readers interested in applying statistical concepts and widely-used mathematical formulas for estimators and their standard-errors. Section 4 discusses special topics required in estimation approaches using dynamic models, including dynamic programming. Sections 5, 6 and 7 discusses contemporary advances in the context of principal-agent theory, disclosure theory, and earnings management, respectively. Section 8 concludes.

1 Two Hands-On Examples

Below, we develop examples that will serve to illustrate the process of writing a structural model and how it can be estimated, using basic intuition only and (almost) no econometrics. The first example is an approach to estimate disclosure costs, from observations about voluntary disclosures. The second example is a dynamic model to recover stickiness in cost structure.

1.1 Estimating disclosure costs

In this first example, we examine a class of problems in which alternative methods are useful to recover economic primitives and which involve minimal use of computational or econometric tools. Drawing from the literature, we illustrate the methods drawing from three approaches from prior literature: a method-of-moments approach (Bertomeu, Beyer and Taylor 2016), a maximum likelihood approach (Bertomeu et al. 2020) and the non-parametric approach of Cheynel and Liu-Watts (2020). We show specifically how a consistent estimator can be derived from theoretical properties of a model and how to numerically compare the properties of the estimation.

Consider a firm contemplating a decision to make an earnings forecast. The firm privately observes future cash flows $\tilde{x}$, drawn from a distribution with p.d.f. $f(.)$ and makes a decision $d(x) \in \{x, NI\}$, where $d(x) = x$ indicates disclosure while $d(x) = NI$ indicates strategic withholding. The firm maximizes its price post disclosure decision but faces a cost $c > 0$ when making a disclosure. Assume that $\tilde{x}$ has mean normalized to zero: for example, one may think about an earnings surprise after netting out an analyst consensus prior to the disclosure.

Denoting $P(d(x))$ as the market price given a disclosure $d(x)$, the firm discloses if $P(x) - c \geq P(ND)$, i.e., its disclosure price is greater than its withholding price. To close the model, suppose that the firm is priced at its expected cash flow so that $P(x) = x$ and $P(ND) = \mathbb{E}(\tilde{x}|d(\tilde{x}) = ND)$ is priced at the expected information that would be withheld.

As is well-known in these models, the disclosure strategy reduces to a threshold τ such that firms disclose when their information $x \geq \tau$ is sufficiently favorable.[2] A firm with information exactly at the threshold must be indifferent between disclosing and with-

[2] See, for example, Jovanovic (1982) or Verrecchia (1983).

holding, implying a condition for the marginal discloser:

$$\tau - c = P(ND) = \mathbb{E}(\tilde{x}|\tilde{x} \leq \tau), \tag{1.1}$$

where the left-hand side is the payoff net of cost from disclosure, while the right-hand side is the payoff from withholding information.

The usual next steps in the context of a theoretical problem is to assume that the parameters of the model are known. The theoretical researcher thus goes from known parameters to predictions about data. A structural model, by contrast, proceeds in reverse. The theoretical parameters are unknown so the objective is to go from known features in the data to usually unobserved theoretical parameters.

There are here two approaches to solving this problem: (1) to solve the model in its entirety and fit the predicted properties of the model to data analogues, and (2) to estimate the model using properties of the model but without solving the game. Of these two methods, (1) is usually the conceptually simpler approach because it involves the same steps as solving theory: if a model can be numerically solved, it can be structurally estimated. However, (2) can be much easier to implement and is usually computationally more economical.

The first approach is directly solve the model: solving the model here amounts to deriving the endogenous optimal threshold τ and withholding price $P(ND)$. To do this, we will need to assume knowledge of the distribution $F(.)$, so let us assume that $F(.)$ is Normally distributed with, as noted earlier, mean zero and variance σ^2. Using Mill's ratio for the mean of a truncated normal, equation (1.1) becomes

$$\tau_n - c_n = -\frac{\phi(\tau_n)}{\Phi(\tau_n)}, \tag{1.2}$$

where Φ (resp. ϕ) is the c.d.f. (resp. p.d.f.) of the standard normal, and $\tau_n \equiv \tau/\sigma$ and $c_n \equiv c/\sigma$ are the standardized threshold and disclosure cost, respectively.

Like many models, the solution is not in closed-form. Verrecchia (1983) shows that (1.2) has a unique solution $\tau_n = T(c_n)$ that can be solved numerically for any c_n. The approach below closely follows the method of moments implementation in Bertomeu et al. (2016). The probability of disclosure predicted by the model is

$$Pr(d(\tilde{x}) = \tilde{x}) = 1 - \Phi(\tau_n) = 1 - \Phi(T(c_n)), \tag{1.3}$$

so, given a frequency of disclosure observed empirically $\hat{p}$, a disclosure cost can be readily estimated by solving numerically for $\hat{c}_n$ solution to

$$\hat{p} = 1 - \Phi(T(\hat{c}_n)). \tag{1.4}$$

The next step is to estimate the unobserved standard deviation σ. Using the observed standard deviation of disclosures would be inconsistent with the assumptions of the model because disclosures are truncated at the disclosure threshold. Nevertheless, one can write the variance of disclosures for a normal truncated at τ:

$$Var(\tilde{x}|\tilde{x} \geq \tau) = \sigma^2(1 + \tau_n \frac{\phi(\tau_n)}{1 - \Phi(\tau_n)} - (\frac{\phi(\tau_n)}{1 - \Phi(\tau_n)})^2), \tag{1.5}$$

which, denoting $\hat{v}^2$ as the observed variance of disclosures, directly yields an estimate $\hat{\sigma}^2$ by solving the above equation:

$$\hat{\sigma}^2 = \frac{\hat{v}^2}{1 + T(\hat{c}_n)\frac{\phi(T(\hat{c}_n))}{\hat{p}} - (\frac{\phi(T(\hat{c}_n))}{\hat{p}})^2}. \tag{1.6}$$

Finalizing the analysis, an estimate for the cost is obtained as

$$\hat{c} = \hat{\sigma}\hat{c}_n. \tag{1.7}$$

This is not the most efficient estimation procedure because only the disclosure fre-

quency is used to estimate the model. Bertomeu et al. (2020) propose an alternative maximum likelihood estimation procedure, which simplified to this setting, involves directly writing the likelihood of the data. Given a sample of disclosures (d_{it}), the likelihood is

$$l(d_{it}; c, \sigma) = 1_{d_{it}=ND}\Phi(T(c/\sigma)) + 1_{d_{it}=x_{it}\geq\sigma T(c/\sigma)}\frac{1}{\sigma}\phi(\frac{x_{it}}{\sigma}), \tag{1.8}$$

where the first component is the probability $F(\tau) = \Phi(\frac{T(c/\sigma)}{\sigma})$ to fall below the threshold when the observation is a non-disclosure and $\frac{1}{\sigma}\phi(\frac{x_{it}}{\sigma})$ is the density if the observation with a disclosure.

For a given dataset of disclosures (d_{it}), the primitives can then be estimated by maximizing the (log) likelihood

$$(\hat{c}, \hat{\sigma}) \in argmax_{c,\sigma} \sum_{i,t} \ln l(d_{it}; \sigma, c). \tag{1.9}$$

A few differences with the method of moments in (1.7) could be viewed as strengths or weaknesses depending on the match between data and model. The likelihood function uses more detail about the problem by assigning a likelihood to each observation. This will benefit efficiency but makes the model more sensitive to its specification. Here, a single anomalous disclosure with $x_{it} < \sigma T(c/\sigma)$ implies a likelihood of zero. For this reason, implementing a likelihood method often requires to model random factors or unobserved heterogeneity, so that such events with probability zero are not implied by the model. To address this, Bertomeu et al. (2020) assume that the cost c is stochastic. By contrast, the effect of unmodelled noise is less dramatic in the method of moments estimator (1.7) because some of the effect of noise may average out.

Maximization of the likelihood is equivalent to choosing moments equal to the first-order condition on the likelihood function (1.8). Hence, an alternative interpretation of maximum likelihood is as choosing moments optimally using statistical theory, given that

likelihood-based models are more efficient than moment estimators. This may be a disadvantage if there are known economic moments that, being key implications of the theory, should be targeted while other implications of the model have been simplified and are not objects of interest. Method of moments thus requires more economic intuition to guide the choice of moments and the researcher may have more parameters to estimate than reasonable moments. The choice of critical incidental moments may also be contaminated by researcher priors about acceptable results. A likelihood-based method takes the process of choosing moments out of the hands of the researcher.

Structural models can accommodate heterogeneity and one can add heterogeneity to a sample by writing $c_{it} = \beta' X_{it}$ where β is a vector to be estimated and X_{it} are firm time-varying characteristics. Under maximum likelihood, the vector β can be estimated in lieu of c and nothing in the estimation needs to be changed. With a method of moments, by contrast, the researcher will be forced to find a number of motivated additional moments equal to the dimension of X_{it}; if there are many characteristics, choosing which moments best capture the model can be challenging.

Solving the model requires to specify knowledge of distributions which, in practice, may not necessarily fit properties of a sample. But not all structural models require distributional assumptions. The approach adopted by Cheynel and Liu-Watts (2020) shows that a disclosure cost can be estimated without solving the model for an optimal threshold $T(.)$. Given that the mean of surprises $\tilde{x}$ is normalized to zero, the law of total expectations implies that:

$$F(\tau)\mathbb{E}(\tilde{x}|\tilde{x} \leq \tau) + (1 - F(\tau))\mathbb{E}(\tilde{x}|\tilde{x} \geq \tau) = \mathbb{E}(\tilde{x}) = 0, \tag{1.10}$$

by averaging the expected cash flow conditional on non-disclosure and conditional on

non-disclosure. Equation (1.10) can then be rearranged into

$$\mathbb{E}(\tilde{x}|\tilde{x} \leq \tau) = -\frac{1 - F(\tau)}{F(\tau)}\mathbb{E}(\tilde{x}|\tilde{x} \geq \tau). \tag{1.11}$$

Reinjecting this expression into the indifference condition (1.1) yields

$$c = \tau - \mathbb{E}(\tilde{x}|\tilde{x} \leq \tau) = \tau + \frac{1 - F(\tau)}{F(\tau)}\mathbb{E}(\tilde{x}|\tilde{x} \geq \tau). \tag{1.12}$$

which, as used in Cheynel and Liu-Watts (2020), yields an estimation procedure

$$\hat{c} = \hat{\tau} + \frac{\hat{p}}{1 - \hat{p}}\hat{e}, \tag{1.13}$$

where the three terms in the right-hand side can be obtained empirically: $\hat{\tau}$ is the minimal disclosure in the sample and consistently estimates the disclosure threshold, $\hat{p}$ is the disclosure frequency and $\hat{e}$ is the average disclosure. Unlike full solution methods, the estimate is in closed-form and requires no distributional assumption about $F(.)$ and, as will be shown when revisiting this estimator in Section 2.4, is asymptotically unbiased even in the presence of uncertainty about information endowment á la Dye (1985).

To examine the properties of these estimators in finite samples, $\tilde{x}$ is drawn from a standard normal distribution with $c = .25$. The model is simulated by solving numerically for $T(.25)$ such that, for each observation $\tilde{x}_{it}$, the observation is censored when $\tilde{x}_{it} \leq T(.25)$. To measure the potential bias, we simulate $1,000$ datasets for various sample sizes, calculate the estimators BBT from Bertomeu et al. (2015) using the moment condition (1.7), BBM from Bertomeu et al. (2020) using the likelihood function (1.9) and CLW from Cheynel and Liu-Watts (2020) using the non-parametric approach in (1.13), averaging over all simulated datasets. A one-standard deviation interval is obtained as the standard-error of the simulated estimators.

While all estimators are consistent, Figure 2 shows that the estimators have low finite

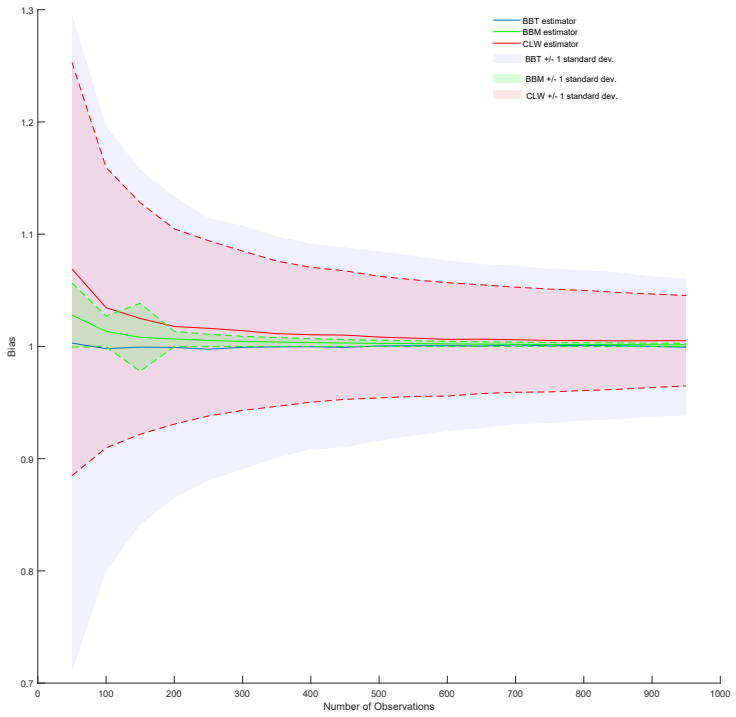


Figure 2: Evaluating the finite-sample bias and efficiency of estimation procedures

sample bias for sample size of 200 or above, i.e., their expectation is close to the true parameter of one. The estimation procedure with the least finite sample bias is the parametric BBT model, followed by BBM and then CLW. However, the methods also differ in efficiency. The smallest confidence interval is attained with BBM, as expected from the efficiency properties of maximum likelihood, followed by CLW and, then, BBT.

1.2 Estimating Asymmetries in Price to Earnings

The next example features a problem that is inherently dynamic and requires the model to frame a sequence of choices over multiple periods. Breuer and Windisch (2019a) consider the problem of understanding asymmetries in the relation between prices and earn-

ings given that in the following Basu (1997) regression:

$$\frac{e_t}{P_{t-1}} = a_0 + a_1 r_t + a_2 r_t 1_{r_t<0}, \tag{1.14}$$

where e_t indicates earnings scaled by lagged price P_{t-1} and r_t indicates returns.

A positive coefficient $a_2 > 0$ is usually interpreted as accounting conservatism because earnings incorporate a greater proportion of the market news conditional on losses. This interpretation has received considerable attention in the literature, as the asymmetry from "Basu coefficients" is commonly interpreted as measuring conservatism, see, e.g., Khan and Watts (2009), Lara, Osma and Penalva (2011), or Bertomeu, Cheynel, Liao and Milone (2021d).[3]

However, in a model where the firm makes optimal investment, the relation between earnings and prices need not be linear. Breuer and Windisch (2019a), hereafter BW, ask whether a benchmark where earnings are unbiased would feature such a relationship. To answer this question, they solve a model in which the firm makes optimal investments and maintains its capital base.

Time is indexed by $t \in [0, \infty)$. Each period, the firm makes an investment choice $I_t \geq 0$, implying a law of motion for the stock of capital

$$k_t = k_{t-1}(1 - \delta) + I_t, \tag{1.15}$$

where $\delta \in (0, 1)$ indicates a rate of depreciation, I_t is investment and k_t is current capital stock. Then, the firm achieves a cash flow

$$CF_t = z_t k_t^{\theta} - \psi \frac{I_t^2}{k_t} - I_t, \tag{1.16}$$

[3]To be fair, the measures are typically used in a relative sense with a higher measure indicating "more" conservatism; a coefficient zero need not mean a neutral reporting system consistent with the analysis of Breuer and Windisch (2019a). The interpretation of asymmetric timeliness as a comparative static is further developed in Ball, Kothari and Nikolaev (2013).

where $\theta \in (0,1)$ is a decreasing returns to scale Cobb-Douglas technology and ψ is the disruption cost due to new investments. The process of innovation $z_t = \rho z_{t-1} + (1-\rho)\underline{z} + \epsilon_t$ is a serially correlated productivity shock and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ is an i.i.d. normally-distributed innovation. As in Breuer and Windisch (2019a), this process is assumed Normal to capture operational losses and the firm cannot disinvest $I_t \geq 0$. Hereafter,

$$f(z_{t+1}|z_t) = \frac{1}{\sigma_\epsilon}\phi\left(\frac{z_{t+1} - \rho z_t - (1-\rho)\underline{z}}{\sigma_\epsilon}\right) \tag{1.17}$$

denotes the p.d.f. of $z_{t+1}|z_t$.

Each period, the firm observes z_t and makes an investment to maximize its discounted cash flows according to the following neo-classical production technology

$$V^*(z_0, k_0) = \max_{(I_t)} \mathbb{E}\left(\sum_{t=0}^{\infty} \beta^t CF_t | z_0, k_0\right) \tag{1.18}$$

discounted at $\beta \in (0,1)$. Because $(I_t)_{t=0}^{\infty}$ is a complex infinite-dimensional vector of all past history of realized productivity shocks, maximizing the right-hand side directly is not practical. The Bellman principle implies that the value function $V^*(k,z)$ can be interpreted as the value achieved by the firm at *any* state (k,z). This important observation motivates the following Bellman equation:

$$V^*(k,z) = \max_{I \geq 0} \{ \underbrace{z k^\theta - \psi \frac{I^2}{k} - I}_{(A)} + \beta \underbrace{\int f(z'|z) V^*((1-\delta)k + I, z') dz'}_{(B)} \}. \tag{1.19}$$

This functional equation states that the value to the firm $V^*(k,z)$ can be decomposed as the current payoff from making a single optimal investment (A) plus the discounted expected value in the following period (B) once the capital is updated to the new state $k' = (1-\delta)k + I$ and the next-period productivity shock z' is drawn.

Standard results guarantee that this type problem, in which a separable bounded smooth

per-period payoffs is discounted, admits a single solution $V^*(k,z)$ that can be calculated by iterating over Equation (1.19) starting from an initial guess (Stokey, Lucas and Prescott 1989). Specifically, starting from a guess $V^0(k,z)$, for example $V^0(k,z)=0$, one can construct a sequence of updated guesses $V^i(k,z)$ by solving

$$V^{i+1}(k,z)=\max_I\{zk^\theta-\psi\frac{I^2}{k}-I+\beta\int f(z'|z)V^i((1-\delta)k+I,z')dz'\}. \quad (1.20)$$

At each "iteration." the researcher uses the current guess $V^i(k,z)$, solves numerically the right-hand side of (1.20) and then uses the maximum to calculate a new updated guess $V^{i+1}(k,z)$. This new guess is then used in the next iteration on the left-hand side to obtain $V^{i+2}(.)$ and the process continues until a norm $||V^{i+1}-V^i||$ is small indicating convergence to $V^*(k,z)$.

Because it is not possible to solve equation (1.20) over a continuous set of states, a common method is to discretize possible values of (k,z) into a finite grid $K\times Z$, and then replace the program by a discrete analogue, for any $(k,z)\in K\times Z$,

$$V^{i+1}(k,z)=\max_{k'\in\hat{K}}\{zk^\theta-\psi\frac{(k'-(1-\delta)k)^2}{k}-(k'-(1-\delta)k)+\beta\sum_{z'\in Z}Pr(z'|z)V^i(k',z')dz'\}, \quad (1.21)$$

where $\hat{K}=K\cap\{k'-(1-\delta)k\geq 0\}$ guarantees that investment remains positive and $Pr(z'|z)$ is defined as a discrete approximation of the $z_{t+1}|z_t$ on the grid Z. For autoregressive Normal processes, the Tauchen (1986) method suggests a grid Z and transition matrix $Pr(z'|z)$ and has code widely available for most programming languages. In the equation above, rephrasing the optimization in terms of the choice of next-period capital k' and rewriting investment as $I=k'-(1-\delta)k$ makes it easier to guarantee that policies are chosen on the grid K. When mapping the model to empirical observable variables, $P_t\equiv V(k_t,z_t)$ represents the price, k_t is the firm's assets and income e_t can be recovered

as

$$e_t \equiv z_t k_t^{\theta} - \gamma \frac{I_t^2}{k} - \delta k_t. \tag{1.22}$$

This model can be estimated by simulated method of moments which involves minimizing the difference between moments in the data and moments predicted by the model. To obtain the latter, the model is first numerically solved and, then, theoretical moments are calculated by simulating a large enough number of data points by sampling a process (z_t) and applying the optimal investment policy from the model numerical solution. In Breuer and Windisch (2019b), this model is estimated by equally weighting differences in mean, standard-deviation and auto-correlations of price-to-book, investment and income.

The analysis below simulates data from the following closely related parameter values $(\overline{z}, \sigma_\epsilon, \rho, \beta, \theta, \psi, \delta) = (1, 1, .5, .9, .33, .5, .3)$. Figure 3 is a non-parametric fit of earnings on returns, and reveals the shape found in Basu (1997) in which earnings flatten with higher returns. High returns are driven by high realized productivity shocks z_t and lead to high investment I_t; but these investment reduce current earnings by I_t^2 which flatten earnings. Earnings are conservative not because current good news are being deferred ("conditional" conservatism) but because investment is being fully expensed while future revenues from these investments are not matched to current earnings. Put differently, BW reinterpret Basu coefficients as unconditional conservatism, caused deductions of certain items (such as investment costs).

This model can be easily estimated and, to illustrate an estimation procedure in the simplest setting, one can use the method of moments where theoretical moments are matched to model moments. The first step of this estimation is to select a number of moments equal to the number of parameters which are likely to contain information about the parameters. For illustration, suppose that all parameters are common-knowledge at the values of Figure 3 except for the returns on capital and the adjustment cost set at

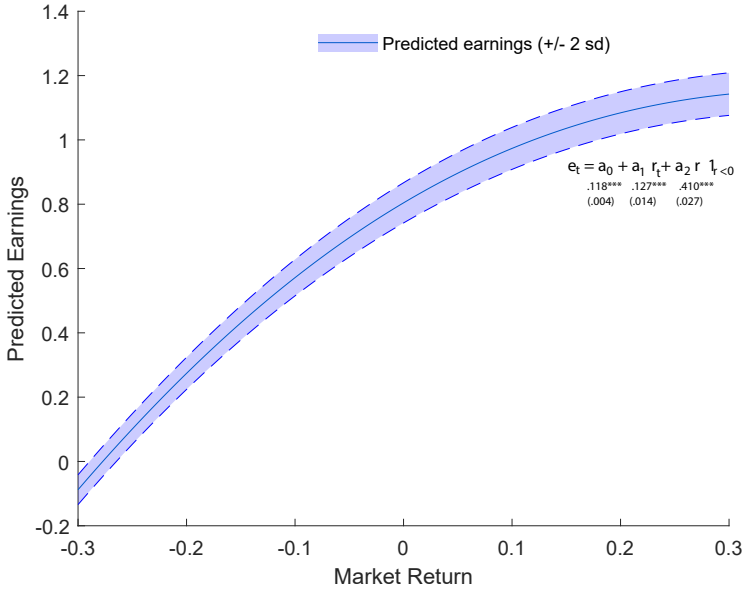


Figure 3: Asymmetry in Earnings-to-Price

$(\theta_0, \psi_0) = (.33, .5)$ but unknown to the econometrician. Since the model explains the asymmetric relation between earnings and returns, let us match the coefficients a_1 and a_2 in equation (1.14). By solving and simulating a large data set from the model (usually 5 to 10 times the size of the sample), the researcher can compute the model-implied coefficients $m(\theta, \psi) = (a(\theta, \psi), b(\theta, \psi))$ at an adequate level of precision. The (simulated) method of moment estimator is then to maximize

$$(\hat{\theta}, \hat{\psi}) \in argmax_{\theta,\psi}(m(\theta, \psi) - m_0)'(m(\theta, \psi) - m_0), \tag{1.23}$$

where m_0 is the moment in the actual data.

For this exercise, we use a simulated dataset from $(\theta_0, \psi_0) = (.33, .5)$ and treat this simulation thereafter as the empirical dataset; of course, in practice, one would use real data. The moments in the empirical sample are $m_0 = (.13, .41)$. Maximizing in (1.23) using a global search algorithm (particleswarm in Matlab) yields parameter estimates $(\hat{\theta}, \hat{\psi}) = (.30, .30)$, with a slight error relative to the true ψ_0. One can also add one or more additional moments to capture other aspects of the model: suppose we add the variance

of investment as an additional moment. In the simulated data, $Var(I_t) = .011$. Adding this moment to the estimation yields a slight improvement to the parameter estimate at $(\hat{\theta}, \hat{\psi}) = (.30, .37)$.

However, a model with more moments than parameters need not be more accurate if the additional moments add noise. Hansen (1982) proposes to uses an optimal weight matrix W and solve the following objective function:

$$(\hat{\theta}, \hat{\psi}) \in argmax_{\theta,\psi}(m(\theta, \psi) - m_0)'W(m(\theta, \psi) - m_0). \tag{1.24}$$

The optimal choice of the weighting matrix can be obtained as the variance of moment conditions, which is a simple process to derive and will be explained in details in Section 2.4. In this example, using the optimal weight matrix yields an improved estimate $(\hat{\theta}, \hat{\psi}) = (.30, .40)$.

2 A Simplified Approach to Structural Estimation

2.1 First Steps

Because structural models are non-linear and may imply complex statistical implications, the econometrics required to estimate a model are often application-specific. Standard asymptotic theory, such as pre-packaged tools in linear models, is not available or needs to be adapted to specifics of the theoretical model, in turn requiring more statistical knowledge than reduced-form analysis. Unfortunately, this added complexity increases the barriers to entry to writing, understanding and interpreting structural models, and can occasionally lead researchers to rule out important economic implications to keep the econometrics simple. This section develops a simplified computational approach allowing researchers to estimate a wide class of models without heavy investment in econometrics or numerical methods.

A structural model starts with the specification of a theoretical model: for our purpose here, we should think about a model as a mapping from parameters θ to a vector real-valued predictions $G(\theta) = (G_1(\theta), \ldots, G_M(\theta))'$. For some applications such as estimating disclosure costs in 1.1, the researcher solves the model in closed-form and writes the function $G(\theta)$ explicitly. From a programming standpoint, if $G(.)$ does not have a closed-form expression, one can solve the model numerically (subprogram 1), simulate a large dataset from the solution (subprogram 2) and compute $G(.)$ from the simulated dataset (subprogram 3). In 1.2, for example, the researcher solves the value function to derive the optimal investment strategy, simulates a panel of firms and computes $G(.)$ by running a Basu regression on the simulated data.[4]

The objective of the structural model is to match predictions of the model $G(.)$ to what is observed empirically. Let an empirical sample be given by $X = (x_i)_{i=1}^n$ and suppose that there exists a function $\hat{G}(X)$, in short $\hat{G}$, that is a consistent estimator of $G(\theta_0)$ if the sample was generated from θ_0. In many applications, the choice of $\hat{G}$ is self-evident from the context. For example, if $G(.)$ was defined as a set of moments, the same moments can be defined on the sample X. When $G(.)$ was computed from a simulation, the function $\hat{G}$ should normally be computed by running the same procedures (e.g., subprogram 3) swapping the simulated data with the real data X. This can offer great flexibility when $G(.)$ involves pre-cleaning of the data or non-parametric steps. For example, one could be winsorizing all observations and running a fixed-effects regression to remove heterogeneity in the data, or, to relax some functional forms, involve a non-parametric estimation of a density or of a conditional expectation. As long as all steps are applied to both data and simulation, the approach will match the same objects.

[4]One would of course want the simulation noise to be as small possible, so it is usually practical to set the simulation size as large as possible so that the resulting $G(.)$ remains practically constant when changing the random number generator (seed) in the simulation code. Usually, the computationally expensive part of this approach is solving the model (subprogram 1) so increasing simulation size is rarely constrained by computing power.

To estimate the parameters of interest, the researcher minimizes an objective function:

$$\hat{\theta} \in \text{argmin}_{\theta} \quad (G(\theta) - \hat{G})' W (G(\theta) - \hat{G}), \tag{2.1}$$

where W is a weighting matrix, for example setting W as the identity implies the sum of squared differences between model and data predictions. In this section, we will assume that the model is locally (point) identified, that is, has a unique solution. A formal discussion of identification is deferred until later sections.

Given that the empirical sample has finite size, the estimation may be inefficient if some dimensions of the function $\hat{G}$ feature more sample noise than others. Intuitively, the researcher should weigh predictions that are accurately computed in the sample. A common method is to weigh elements of $\hat{G}$ according to the inverse of the covariance matrix $W = Var(\hat{G})^{-1}$.[5]

In principle, because the model is always assumed to be true when conducting an estimation, this variance can be computed by computing $G(\hat{\theta})$ many times from simulated data of the same size as the sample and for different random seeds (subprogram 2), and computing the covariance matrix of the resulting vectors $G(\theta)$ (subprogram 3). However, in practice, this method is unnecessary and can magnify the effect of specification errors when, as is common, the model is simpler (has fewer error terms) than the data-generating process. Instead, a more direct approach is to resample (with replacement) from the data X to form a large number $s = 1, \dots S$ of resampled datasets $X_i^s = (x_i^s)_{i=1}^n$ and compute the resulting $\hat{G}$ for each X_i^s. This alternative method does not require the assumption that

[5]The choice of weighting moments draws a non-trivial philosophical question: when are weights necessary given that (2.1) with the identity matrix is a consistent and easier to interpret criterion? Many studies use the identity matrix or a simplified approach $Diag(Var(\hat{G}_1)^{-1}, \dots, Var(\hat{G}_I)^{-1})$, especially if the resulting standard-errors from these simplified procedures are already sufficiently small so that reducing them further would not affect the economic inference. Having noted this, *not* weighting can make the estimation very sensitive to the inclusion of noisy predictions and, if the researcher has discretion over which moment to use and how to scale them, may make the general inference less credible. For this reason, it is usually preferable to use a weighting scheme consistently unless the $G(.)$ and $\hat{G}$ are uniquely suggested by the applied context.

the model is correct to derive the weights and is also far simpler to implement because it does not require knowledge of $\hat{\theta}$ or solving the model, nor does it carry over estimation noise in $\hat{\theta}$. One needs only write a program that randomly samples observations from the original dataset X (subprogram 4).

Standard errors and some test statistics can be obtained using a similar procedure, as long as the model is quick to solve numerically. To recover standard-errors, one can reestimate (2.1) after swapping the resampled X^s instead of the original X, to obtain the empirical distribution of the estimator from $(\hat{\theta}_s)_{s=1}^S$. Standard-errors can then be obtained by computing the covariance matrix of $(\hat{\theta}_s)_{s=1}^S$, and taking the square root of the diagonal term. A Student-t like hypothesis testing can be also be performed from the empirical distribution. Suppose that the researcher wishes to test whether "$g(\theta_0) = 0$" where $g(.)$ is a set of restrictions with a specific economic interpretation. For example, the researchers may wish to know if several parameters of θ_0 specific to a certain mechanism are zero. A Wald-like test statistic is given by

$$W \equiv ng(\hat{\theta})'(g'(\hat{\theta})'Var(\hat{\theta})g'(\theta))^{-1}g(\hat{\theta}), \tag{2.2}$$

where g' indicates the gradient of the restriction being tested. Under commonly-satisfied regularity conditions, this test statistic converges to a χ^2 distribution with degrees of freedom which is usually the number of dimensions of $g(.)$. However, the test can also be performed, usually with better finite-sample properties, by computing the empirical distribution of (W^s), where each W^s is the resampled test statistic in (2.2) using $\hat{\theta}_s$. Then, a p-value is the fraction of all W^s such that $W \geq W^s$.[6]

We conclude on the use of simulations for counter-factual analysis. The most common presentation of counter-factuals is in the form of implications at the estimated parameter

[6]For pivotal test statistics, that is, their asymptotic distribution does not vary in the parameters, the empirical distribution has better finite-sample properties than the asymptotic distribution (Horowitz 2001). However, this does not mean that the method performs worse than asymptotic theory when a statistic is not pivotal and, often, different approaches perform to a sufficient degree of accuracy for common hypothesis testing.

$\hat{\theta}$ and does not require re-sampling. A more complete approach may involve confidence intervals on counter-factuals given that $\hat{\theta}$ is not precisely known. This can be done using the same methods as those above, simulating the counter-factuals of interest for $\hat{\theta}_s$ and using the results to compute standard-errors or confidence intervals.

In conclusion, the method of re-sampling is versatile, and can be used throughout to obtain weights, standard-errors on estimated parameters, test statistics and standard errors on counter-factuals. Its primary computational cost is that it requires re-estimation of $\hat{\theta}_s$ which increases the estimation by a factor of S. This is only possible if the model is sufficiently quick to simulate - nevertheless, because the re-estimations are separately done, the process can be easily parallelized. Further, a notable computational step is the initial set-up of the optimization (2.1) because the optimization routine needs to know the space of the global search requiring trial and error. For a model that is reasonably estimated, estimates should be stable around $\hat{\theta}$ and, therefore, the set-up used for the actual sample should yield valid interior optima in samples X^s.

2.2 Moment Conditions

Next, we present the theory discussed in 2.1 with greater formalism but leaving for now to Section 3 a complete discussion of identification and asymptotics of structural models. To this effect, we now specialize the function $G(.)$ to moment conditions, which forms the basis of many estimation problems.

The object of interest in an econometric model is a family of distributions $(F(x;\theta))_{\theta\in\Theta}$ of an empirically observable random variable $\tilde{x}$, where θ is a parameter unknown to the researcher with true value θ_0. The researcher knows a vector of M model-implied restriction $G_0(.)$ such that $\theta = \theta_0$ if and only if

$$G_0(F(x;\theta)) = \mathbf{0}. \tag{2.3}$$

$$M \times 1 \qquad M \times 1 \tag{2.4}$$

Because $F(x;\theta_0)$ is not directly observable, one cannot directly solve (2.3) as a system of equations to recover θ_0. Suppose that the researcher observes a dataset $(x_i)_{i=1}^n$ defined as n draws from $F(x;\theta_0)$ such that there exists a sample approximation $G_n((x_i)_{i=1}^n;\theta)$ that converges to $G_0(F(x;\theta))$ as sample size becomes large.[7] In short-hand, we shall use hereafter (x_i), G_n and G_0.

The researcher can then obtain an estimate $\hat{\theta}$ by minimizing the following quadratic objective function

$$\begin{array}{ccccc} \hat{\theta} \in \arg\min_{\theta} Q_n \equiv & G_n' & W & G_n. \\ & 1\times 1 & 1\times M & M\times M & M\times 1 \end{array} \tag{2.5}$$

As we are using a sample approximation of G_0, the restrictions need not be true at equality: when $G_n \neq \mathbf{0}$ for any θ, a weight matrix W assigns which restrictions in G_n are most important. For expositional purposes, we shall initially assume that the researcher has chosen a weight matrix W and discuss later on empirical methods to select W to increase the quality of the estimation.

Example 1: Suppose that $\tilde{x} \sim N(\theta, 1)$ is a family of Normally-distributed random variables with unknown true mean $\theta_0 \in \mathbb{R}$. A natural theoretical restriction for this model G_0 and its sample approximation can be written, respectively, as

$$G_0 \equiv \int x dF(x;\theta) - \theta_0, \qquad G_n \equiv \frac{1}{n}\sum x_i - \theta \tag{2.6}$$

implying an estimate $\hat{\theta} = \frac{1}{n}\sum x_i$ equal to the sample mean.

Example 2: Suppose that $F(x;\theta)$ has a p.d.f $f(x;\theta)$ that can be written in closed-form. Standard results in maximum likelihood estimation suggest that the true parameter

[7] The concept of convergence required to ensure consistency will be discussed in the next section.

must maximize the expected log likelihood which, writing the first-order condition of this maximization, implies the following restriction:

$$G_0 \equiv \int \frac{\partial \ln f(x;\theta)}{\partial \theta}|_{\theta=\theta_0} dF(x,\theta_0) = 0. \tag{2.7}$$

The sample analogue to the expectation above is then defined as

$$G_n \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta}. \tag{2.8}$$

When implementing maximum likelihood, most researchers prefer to directly maximize $\frac{1}{n}\sum_{i=1}^{n} \ln f(x_i;\theta)$ over minimizing its first-order conditions G_n given the two approaches are in principle equivalent.

Example 3: Suppose that the researcher knows M theoretical moments which, according to the model, should be equal to $\phi(\theta)$ which can be either calculated analytically (as in section 1.1) or simulated after solving the model numerically (as in section 1.2). Therefore, the restriction can be written

$$G_0 = H(F(x;\theta)) - \phi(\theta_0), \tag{2.9}$$

where $H(.)$ is the equation of the moment. As an example, $H(F) \equiv \int x dF(x)$ if the moment is a mean or $H(F) \equiv \int x^2 dF(x) - (\int x dF(x))^2$ if the moment is a variance and, more generally, $H(.)$ can represent any function of $F(x;\theta)$. What matters here is that the moment can be approximated by a sample analogue H_n (the same moment calculated with the sample) and define

$$G_n \equiv H_n - \phi(\theta), \tag{2.10}$$

by calculating the difference between sample and theoretical moment. The moment estimator is then obtained by minimizing $Q_n \equiv (H_n - \phi(\theta))'W(H_n - \phi(\theta))$.

2.3 Introduction to the Bootstrap

The methods discussed in 2.1 are applications of the bootstrap, defined as a set of tools such that properties of statistics under consideration are obtained by resampling to form an empirical distribution. The bootstrap method was proposed by Efron (1979) and since, then, has been the object of numerous applications in statistics, economics and forms the core of tuning multi-step machine learning algorithms where deriving closed-form asymptotics is impractical.

The intuition for the method is to construct an empirical c.d.f. $F_n(x)$. which can be directly recovered from a dataset (x_i), to infer properties of statistics constructed from the dataset. As the dataset becomes large, this empirical c.d.f. approximates the true c.d.f. $F(.)$ so that, under certain regularity conditions, one expects that functions of the empirical c.d.f. should have properties similar to functions of the true distribution. Practically, the researcher draws randomly several samples from $F_n(.)$, usually by drawing observations from the empirical sample, calculate the statistic of interest in each bootstrap sample and, for example, estimate its variance as the variance of all bootstrapped statistics.

The bootstrap has been shown to consistently estimate the distribution of statistics that are asymptotically linear and asymptotically normal which, under usual regularity condition (Hayashi 2000), include common extremum estimators such as maximum likelihood or generalized method of moments (Mammen 1992, Mammen 2012, Horowitz 2019). If a statistic is pivotal, that is, its asymptotic distribution does not depend on the parameter θ, bootstrap estimates are more precise than traditional formulas from asymptotic theory. Hence, a textbook recommendation is, *when possible*, to choose statistics that depend less on parameters (Horowitz 2001, MacKinnon 2006). These asymptotic improvements can often be applied to bootstrapping test statistics, such as a chi-square test. However, be-

ing pivotal is not a necessary condition for consistency of the bootstrap and many useful applications of the bootstrap, such as finding the standard-error of a parameter estimate, may not lead to a pivotal statistic and yet may lead to more precise estimates.

There are two benefits from the bootstrap that are unrelated to its potential efficiency gains. First, the bootstrap allows a researcher to derive asymptotic variances using generic programs that do not require deep knowledge of the econometrics of a model. This can be particularly useful if the model is not standard or is not in closed-form. Second, most analytical methods require computation of numerical derivatives to derive asymptotic variances, e.g., gradient of moments or information matrix. Choosing the right step for numerical derivatives can be challenging when other aspects of the model, such as constructing a grid to solve the model, non-trivially interact with the step size. The bootstrap, by contrast, need not require differentiation.

Consider first a simplified problem in which we are interested in some statistical property of a statistic S_n: for example, we seek the variance of parameter estimates. In principle, having estimated $\hat{\theta}$, one could derive the distribution of S_n. However, the last step may be non-trivial if the distributions $F(x;\theta)$ were not fully specified and/or the expression of S_n is non-linear and not in closed-form. To give an extreme example, the researcher may have plugged in a relationship estimated in first-stage from an ensemble learning method such as gradient boosted trees (Friedman 2002) for which there exists no closed-form expression of standard errors.

The bootstrap is a method using the empirical distribution of the dataset (x_i), hereafter $F_n(x)$, to estimate the true distribution $F(x;\theta_0)$. Denoting $\overline{F}(.)$ as the distribution of the entire sample $(\tilde{x}_i)$, one can use a sample approximation $\overline{F}_n(.)$ by randomly drawing n observations with replacement to form a bootstrap dataset (x_i^k). In turn, one can then approximate the true distribution $F^{S_n}(S;\theta_0)$ of the statistic S_n, by drawing from the empirical distribution $F_b^{S_n}(.)$ implied by the bootstrap samples, i.e., the distribution of (S_n^k) where S_n^k is the statistic computed in the bootstrap sample (x_i^k).

Having a procedure to sample from $F_b^{S_n}(.)$, suppose that the researcher wishes to recover a property $H_0(F^{S_n})$; then, the bootstrap suggests to use the following estimator

$$\hat{H}_0(F^{S_n}) = H((S_n^k)_{k\in[1,K]}), \tag{2.11}$$

where $H(.)$ is a function of the sample that approximates $\hat{H}_0$. As n increases, the empirical distribution $\overline{F}^n$ converges to $\overline{F}$ so that, if H_0 and S_n^k satisfy standard smoothness properties, $\hat{H}_0(F^{S_n})$ will converge to $H_0(F^{S_n})$. To obtain the right hand-side of (2.11), the researcher needs to draw bootstrap samples and be able to compute S_n^k but does not need to derive analytically a statistical property of S_n.

Example 4: Let $S_n = \hat{\theta}$ be a parameter estimate obtained by minimizing Q_n in (2.5) and suppose that we are interested in estimating (1) the finite-sample bias and (2) the variance of this estimate. The bias $B = \mathbb{E}(\hat{\theta}) - \theta_0$ can be estimated by

$$\hat{B} = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\hat{\theta}^k}_{\equiv m_1} - \hat{\theta}, \tag{2.12}$$

as the mean bootstrap estimate minus the base estimate. The covariance $V = Var(\hat{\theta})$ is similarly obtained as the sample covariance of the bootstrap estimates

$$\hat{V} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}^k - m_1)(\hat{\theta}^k - m_1)'. \tag{2.13}$$

Example 5: Suppose that $\theta = (\theta^1, \ldots, \theta^J)$ is a vector of parameters with true value $\theta_0 = (\theta_0^1, \ldots, \theta_0^J)$. The researcher is interested in examining the prediction of a model in a counter-factual θ_c which differs from θ_0 at $\theta_c^1 \neq \theta_0^1$ but otherwise with $\theta_c^j \neq \theta_0^j$ for any $j \geq 2$. Let us write a quantity relevant to the policy-maker $\zeta(\theta)$ where $\zeta(\theta_0)$ may not be observable in the data. This quantity can be estimated by solving the model at $\hat{\theta}$, that is,

$\zeta(\hat{\theta})$ and, similarly, a counter-factual can be obtained as $\zeta(\hat{\theta}_c)$ using $\hat{\theta}_c = (\theta_c^1, \hat{\theta}^2, \ldots, \hat{\theta}^J)$. Because $\hat{\theta}$ is a noisy estimate of θ_0, the bootstrap can be used to obtain standard-errors on the $\zeta(\hat{\theta})$ and $\zeta(\hat{\theta}_c)$, by computing these quantities in each bootstrap sample and calculating the variance of the vector as in (2.13).

2.4 Optimal Weight Matrix

We have for now taken the weight matrix W as a given. For example, using an $M \times M$ identity matrix minimizes the sum of the squared deviations from the model equations $Q_n = G_n' G_n$. Unfortunately, using the identity is often inefficient because not all equations in the vector G_n estimate θ precisely and, further, the scaling of each equation can affect the estimation by implicitly putting more weight on dimensions with greater scale. Practically, an identity weight matrix can make the estimation sensitive to adding relations that are not relevant for the estimation or are noisily estimated.

Hansen (1982) shows that an optimal weight matrix, in the sense of reducing the asymptotic variance of the estimator, can be obtained as $W = (Var(G_n))^{-1}$. Intuitively, components of G_n that vary less are closest to the true G_0 and be given more weight in the estimation. There are standard analytical methods to recover $Var(G_n)$ when G_n has standard forms, such as being a centered moment or the parameter of a regression. However, problems are no longer standard when using more complex multi-step procedures which may involve, for example, pre-cleaning a dataset through a first-stage fixed effect regression, removing outliers, machine learning or non-parametrics or using variables estimated from another dataset.

The bootstrap can be adapted to any such problem where analytical expressions do not exist or are non-trivial, by setting the statistic of interest as $S_n = G_n$. Then, one can calculate G_n^k in each bootstrap sample and recover an estimated bootstrap weight matrix $\hat{W} = (Var(G_n))^{-1}$. Often, G_n^k is additively separable in θ so that computing a weight matrix does not require knowledge of the true θ_0. If, on the other hand, $G_n(.)$

depends on θ, it is possible to use a consistent estimate $\hat{\theta}$ (for example, using a first-stage identity weight matrix or repeatedly updating the weight matrix for greater finite-sample performance).

Example 3 (cont.): Let the model restrictions be defined as in (2.9) with $G_n = H_n - \phi(\theta)$ equal to the difference between a sample moment H_n and its theoretical model value $\phi(\theta)$. Because this equation is separable in θ, implying that $W = (Var(G_n))^{-1} = (Var(H_n))^{-1}$ can be estimated by using the covariance matrix of H_n^k calculated in each bootstrap sample.

Using the optimal weight matrix provides other benefits such as simplifying the computation of useful test statistics. Suppose that θ is a vector of dimension r and we are interested in testing a Null hypothesis that $\theta_0 \in \Theta_c$, where Θ_c is a compact subset of Θ with dimension r'. The distance test statistic is defined as

$$D = n(Q_n(\hat{\theta}_c) - Q_n(\hat{\theta})), \tag{2.14}$$

where $\hat{\theta}_c = \arg\min_{\theta \in \Theta_c} Q_n(\theta)$ is defined as the parameter estimate on Θ_c. If the model restrictions are appropriately defined so that $\sqrt{n}G_n$ is asymptotically normally distributed, the D statistic converges to a $\chi^2(r - r')$ distribution under the Null.

If the dimension of the parameter space r is less than the number of restrictions M, so that $Q_n(\hat{\theta})$ is typically different from zero, the J test statistic can be used to test whether a large Q_n suggests that the conditions of the model are unlikely to be satisfied empirically. This test examines the specification of the model under the Null hypothesis that $G_0 = \mathbf{0}$. The J test statistic

$$J = nQ_n(\hat{\theta}) \tag{2.15}$$

converges to a $\chi^2(r - r')$ distribution under the Null.

2.5 Parametric Bootstrap

The bootstrap discussed until this point is the version presented by Efron (1979) and requires to draw bootstrap samples *from the dataset*. For this reason, it is sometimes denoted "non-parametric" bootstrap because no functional assumption is placed on the distribution. The parametric bootstrap is a different procedure where, instead of drawing from the empirical distribution of (x_i), the researcher draws each bootstrap sample from $F(.|\hat{\theta})$ by simulating the model. The non-parametric bootstrap is usually preferable because the researcher does not need to know $F(.)$ and, therefore, the procedure is less sensitive to model specification. However, there are certain cases where parametric bootstrap is unavoidable. Below, we discuss four important contexts where parametric bootstrap might be necessary.

First, suppose that the researcher is interested in assessing the theoretical performance of a model "in the lab" over different parameter values. For example, the researcher may want to compare whether an estimation approach is better than another, as in (1.1), or check whether a measure obtained in the model correctly captures certain behaviors. To answer this question, the researcher can construct many parametric bootstrap simulations from the true model and, then, compute the object of interest. This approach is commonly-used in econometrics to assess the finite-sample properties of estimators, and is used in finance and accounting to compare the quality of various procedures, see, e.g., Cheynel and Liu-Watts (2015), Bazdresch, Kahn and Whited (2018), Breuer and Schütt (2019) or Bertomeu et al. (2021d). Parametric bootstrap is unavoidable in these circumstances because the researcher must know from which true parameter values the data is generated, including potentially unobservable variables, in order to measure quantities of interest. Note that parametric bootstrap is used here to answer an econometric "theoretical" question, namely, a question about the performance of an estimation procedure, not an applied question about measurements in a sample.

A second context in which parametric bootstrap is useful is when there are two-way

correlations which cannot be ruled out. For example, in a panel, observations may be correlated within a firm or individual and within a year. Block bootstrap, that is drawing firms instead of observations, is no longer sufficient in this case although there exists sampling procedures that will capture some correlations. A systematic approach to this problem is to directly simulate a dataset from the estimated correlation structure.

Third, non-parametric bootstrap cannot be used to derive the distribution of test statistics because one cannot sample from the Null. For this reason, the D test statistics and J test statistics were assumed to follow (asymptotically) a χ^2 distribution over bootstrapping their distribution. Put differently, one would (almost) never reject the Null if the distribution of these test statistics were bootstrapped from the data. On the other hand, the parametric bootstrap can be used to bootstrap any distribution of a test statistic and improve the quality of this distribution in finite samples. In the case of the D statistic in (2.14), the researcher draws datasets from $F(.|\hat{\theta}_c)$ and $F(.|\hat{\theta})$, and computes bootstrap estimates D^k which can be used to form a confidence interval for D in the sample. Similarly, the empirical distribution of the J test statistics in (2.15) can be obtained by computing J^k in the bootstrap samples drawn from $F(.|\hat{\theta})$.

Fourth, the non-parametric bootstrap is not appropriate for very small samples, because resampling with replacement will be more likely to draw the same observation repeatedly and create spurious correlation in the sample. Unfortunately, small samples would also invalidate standard-errors from asymptotic theory, so the researcher must in this case rely more heavily on the assumptions of the model. For example, a small sample of international conflicts or rare diseases may yield noisy estimates but one would want to reliably measure that noise to evaluate possible policies.

2.6 Limitations and Caveats of the Bootstrap

While the bootstrap often works as a skeleton key to handle problems that would be very difficult using standard analytical methods, there are a few important situations

where it may not be suitable or need not perform better than other methods.

The theory of the bootstrap suggests that bootstrap estimates often have better finite-sample properties than many analytical methods; however, these efficiency gains are mathematically proved in the context of pivotal statistics (Horowitz 2001). Pivotal statistics are statistics whose asymptotic distribution does not depend on the parameter θ; most constructs of interest in applied problems are not pivotal, including estimates or coefficients in a weight matrix. While the bootstrap remains a consistent estimation procedure absent pivotal statistics, it may or may not be more efficient than other methods. This is why Horowitz (2001) recommends bootstrapping pivotal statistics when possible.

There are nevertheless some situations in which using the bootstrap should be done with caution. This can fall into two categories: practical implementation and theoretical issues. Let us begin with the practical implementation:

1. *Excessive computational burden.* Applications of the bootstrap that involve estimating $\hat{\theta}^k$ in each bootstrap sample will require significantly more computational power than analytical methods and will be several orders of magnitudes slower than the baseline estimation. Practically, however, this issue has become less of a concern over time given the increase in processing speed and the fact that estimation in bootstrap samples can be easily allocated to separate CPUs or, even, can run on separate work machines without need for organized parallel computing. In most applications, the computing time is in the process of finding a reasonable baseline model and its relevant model restrictions (by estimating different models) or improving the quality of the computational tools to solve a model - all of this pre-analysis can be used when estimating bootstrap without need for new calculations.[8]

[8] An example of this challenge can nevertheless be found in the related area of machine learning, where the complexity of multi-step prediction models (especially with ensemble learning algorithms) make methods other than the bootstrap infeasible, with recent examples in accounting by Bao, Ke, Li, Yu and Zhang (2020) and Bertomeu, Cheynel, Floyd and Pan (2021c). The algorithms require as input a number of hyperparameters whose optimal choice depends on the data; in our setting, the weight matrix would be an example of hyperparameter except that there is explicit econometric guidance on its choice. Unfortunately,

2. *Managed Code.* Many computational methods cannot be easily automated to run on any sample and researchers will often hand-manage numerical methods, usually by running diagnostics after a solution is found to guarantee that tuning parameters in the methods are appropriate. In the context of baseline estimates, one may (1) check for other global optima or whether the estimate lies on a corner of the search space, (2) check whether a grid used to solve a value function is sufficient or (3) check whether a simulation is sufficiently accurate. However, it is infeasible to manage the code for hundreds of bootstrap samples and, without it, the bootstrap estimates may be of lower quality. For example, searching locally around the baseline estimates in the bootstrap samples may find local optima close to the baseline estimates - suggesting low parameter standard errors - even though that real standard-errors may be much greater.

   This issue implies that estimation code used for bootstrapping must be made more robust to sampling variation, especially if the dataset is small such that subsampling can lead to very different datasets. The estimation code should be robust not solely around the parameter estimates but for variation in the bootstrap samples. One approach is to increase the precision of the numerical methods when bootstrapping or encoding some diagnostics detecting one of the problems (1)-(3).

3. *Clustered Data.* The notations developed so far assume that observations x_i are independently drawn from $F(x;\theta_0)$; however, many datasets feature panels with correlated observations from the same firm ("clusters"). The bootstrap can be adapted to (one-way) clusters by redefining x_i as a vector of all observations in a cluster, e.g., all observations of the same firm or individual in the panel. In this case, bootstrap

certain computations may take weeks to complete so that tuning hyperparameters using many bootstrap repetitions would be very costly. Instead, computer scientists split the sample and "bootstrap" the quality of the predictions with somewhere between 5 and 10 subsamples which would be insufficient to approximate the empirical distribution of prediction quality. To be noted, however, their focus is not on finding or interpreting the *optimal* hyperparameters, but to improve the estimation relative to default hyperparameters.

sampling is known as "block" bootstrap and should be performed by drawing clusters. In practice, when bootstrapping a dataset with different firms, the researcher would randomly draw firms. The number of observations n will be defined as the number of clusters.

4. *Small dataset.* If the dataset is too small (in particular, with clustered data having few clusters), the same observation is likely to be repeated multiple times in the bootstrap samples, which will create correlation between observations of the bootstrapped sample. For this reason, bootstrap is not always a solution to improve estimates given a small sample sizes. If the dataset is too small, the only available solution is to rely more on the model and will require the parametric bootstrap, as discussed earlier.

A second set of problems relate to cases where, in theory, the bootstrap does not correctly approximate the object of interest, as econometricians have looked for pathological counter-examples where the bootstrap fails. To see common features of these settings, note that the bootstrap is an asymptotic theory that relies on the assumption that (1) the empirical distribution of x_i can be used to approximate $F(.|\theta_0)$ and, then, (2) approximates the empirical distribution of a construct of interest S_n. Either property will fail in problems where the dataset does not approximate well the true distribution or the construct is not sufficiently smooth for the approximation to apply. Known cases in which asymptotic variances cannot be obtained by bootstrap include estimators in which does not satisfy a $\sqrt{n}$ asymptotic linear expansion, such as when the parameter is in the boundary of the parameter space, when the estimator converges quicker than $\sqrt{n}$ or matching estimators because the same bootstrap observation can be matched to itself in a bootstrap sample. These exceptions are discussed in Andrews (2000), Abadie and Imbens (2008) and Horowitz (2019). A more complete discussion of conditions to guarantee consistency of bootstrap estimates is given in Horowitz (2001) and Mammen (2012).

3 Econometric Methods in Structural Estimation

3.1 Identification

This section presents a formal treatment of the econometric model and of the concepts of identification and empirical content. The econometric model reflects both researchers' a-priori knowledge about the problem as well as a statement about what is observable to the researcher. An econometric model is a set of distributions for the (possibly vector-valued) random variables $\tilde{x}$ and $\tilde{y}$, where $\tilde{x}$ is latent or empirically unobservable variable while $\tilde{y}$ can be observed. Formally, let the econometric model be a family of distributions $(F(x, y; \theta))_{\theta \in \Theta}$. With a slight abuse in language, let us then define $G(y; \theta)$ as the distribution of the observable $\tilde{y}$ given a parameter θ.

Identification refers to the process of finding the parameter θ_0 that generated a large enough dataset of realizations $Y = (y_i)_{i=1}^n$ of $\tilde{y}$. The parameter is identified if a researcher observing only the observable component of the data-generating process can recover the parameter of the model.

Definition 1 *A parameter θ_0 is point identified if, for any $G(y; \theta_1) = G(y; \theta_0)$, $\theta_1 = \theta_0$. Vice-versa, $\theta_1 \neq \theta_0$ are observationally-equivalent if $G(y; \theta_1) = G(y; \theta_0)$.*

To summarize, a model is identified at θ_0 if all other parameters would lead to distinct observables. When two parameters lead to the same observables, we say that they cannot be distinguished by observation alone (they are observationally-equivalent). Observational equivalence can present a significant challenge because two parameters leading to the same observables can have very different interpretations or implications about optimal policy. In certain settings, an econometric model may be point identified for some parameter values and not for others, if identification holds for a subset of $\theta_0 \in \Theta_I$.

While "endogeneity" is commonly diagnosed in a reduced-form statistical model as a failure of identification, it is not equivalent to the more primitive concept of identifi-

cation (Kahn and Whited 2018). Most observables in a structural model are endogenous because they are consequences of optimal choices: economic restrictions on these endogenous variables may identify a parameter of interest even though exogenous shocks are not directly observable. Vice-versa, an exogenous shock may be unable to identify the parameter if it affects multiple parameters at once, changes the game or does not reflect a decision of interest (Chemla and Hennessy 2021a, Chemla and Hennessy 2021b).

To illustrate an identification problem, consider the following textbook example. The wage of an individual $\tilde{w}$ depends on both an inherent i.i.d. skill $\tilde{s}$ which is unobservable and education $\tilde{e}$, and education depends on inherent skill as well as an i.i.d. individual characteristic $\tilde{\epsilon}$. This implies the following structural equations:

$$\tilde{w} = \beta_s \tilde{s} + \beta_e \tilde{e} \tag{3.1}$$

$$\tilde{e} = \alpha \tilde{s} + \tilde{\epsilon}. \tag{3.2}$$

In this model, the set of observables is $\tilde{y} = (\tilde{w}, \tilde{e})$, while the unobservable is $\tilde{x} = \tilde{s}$. The parameters of the model are $(\beta_s, \beta_e, \alpha)$. Rewriting (3.1)-(3.2) in terms of observables only, the model simplifies to

$$\tilde{w} = (\frac{\beta_s}{\alpha} + \beta_e)\tilde{e}, \tag{3.3}$$

which implies that one can identify $\frac{\beta_s}{\alpha} + \beta_e$ but not the causal effect β_e of education on wages in (3.1). In this setting, the endogenous nature of education $\tilde{e}$ implies that $(\frac{\beta_s}{\alpha} + \beta_e)$ is identified but its subcomponents are not.

The most common method to establish identification is to find a function or estimator that can estimate the true parameters of interest.

Theorem 1 *Let there be a function $\Psi(.)$ such that $\Psi(G(y|\theta)) = \theta$ for all parameters θ, then θ is identified on Θ.*

This theorem is of particular interest in applications where there exists an estimation

procedure for the parameter of interest. If this estimator is consistently estimating θ, it can serve as both proof of identification and a method to empirically recover an estimate.

A model may be so general that, while it is identified, can rationalize any empirical observations, thus being inherently unfalsifiable.[9] Following Heckman and Honore (1990), being falsifiable is described as having "empirical content" with the condition that a model with empirical content should be incompatible, regardless of parameter values, with some (credible) empirical samples.

Definition 2 *An econometric model has empirical content if there exists a distribution of $G(y)$ such that $G(y) \neq F(y|\theta)$ for any θ.*

To show that a model has empirical content, it is sufficient to show that there is a distribution $G(y)$ that would not satisfy the assumptions of the model. In an applied setting, it is usually preferable to think about empirical content for distributions that are not a-priori implausible. While having empirical content is defined as a binary characteristic, the quality of the empirical content falls on a spectrum as a function of the number and plausibility of potential distributions $G(y)$ incompatible with the theory.

[9] In The Logic of Scientific Discovery, Karl Popper criticizes Marxism and psychoanalysis as unalsifiable:

> It was during the summer of 1919 that I began to feel more and more dissatisfied with these three theories the Marxist theory of history, psychoanalysis, and individual psychology; and I began to feel dubious about their claims to scientific status. My problem perhaps first took the simple form, What is wrong with Marxism, psycho-analysis, and individual psychology? Why are they so different from physical theories, from Newtons theory, and especially from the theory of relativity?
>
> I found that those of my friends who were admirers of Marx, Freud, and Adler, were impressed by a number of points common to these theories, and especially by their apparent explanatory power. These theories appeared to be able to explain practically everything that happened within the fields to which they referred. The study of any of them seemed to have the effect of an intellectual conversion or revelation, opening your eyes to a new truth hidden from those not yet initiated. Once your eyes were thus opened you saw confirming instances everywhere: the world was full of verifications of the theory.
>
> Whatever happened always confirmed it. Thus its truth appeared manifest; and unbelievers were clearly people who did not want to see the manifest truth; who refused to see it, either because it was against their class interest, or because of their repressions which were still un-analysed and crying aloud for treatment.

We discuss next an approach used in the literature to address identification problems and which needs not involve more assumptions about the economic problem or better data. The first approach is feasible when the researcher does not need full knowledge of θ_0 to answer the question. For a function $\phi(.)$, let us define $\phi(\theta)$ as a feature of the model.

Definition 3 *A feature $\phi(\theta_0)$ is identified if, for any $G(y;\theta_1) = G(y;\theta_0)$, $\phi(\theta_1) = \phi(\theta_0)$.*

A special case of the use of features is set identification, in which a parameter can be identified to be in a certain range $\theta \in [\theta_1, \theta_2]$. If the range is not much larger than the uncertainty due to estimation noise, a set identified parameter may be practically as useful as a point estimate. For a model that is not identified, one can redefine the inference problem to identify a feature $\phi(\theta) = \{\theta' : G(y;\theta) = G(y;\theta')\}$ to represent the equivalent class of all observationally equivalent parameters.

A different approach to establish identification is by using priors and forming an approach to estimation known as a Bayesian statistics. In a Bayesian model, θ is the realization of a known distribution for $\tilde{\theta}$, usually (but not necessarily) a uniform distribution if the parameter space is bounded or a Normal is the parameter is unbounded - these two distributions model maximal ignorance by setting the prior to greatest entropy. The inference process is then given by the posterior distribution $\tilde{\theta}|\tilde{y}$ which is now jointly informed by the prior and the observable. Note that, if θ is identified, this will lead to complete knowledge of the realized $\tilde{\theta}$ regardless of the prior for a large enough sample size.

3.2 Extremum Estimators

The most common types of estimation procedures used in the social sciences are extremum estimators. A formal but accessible treatment of this type of estimators can be found in Hayashi (2000), chapter 7.

Definition 4 *An estimator $\hat{\theta}$ is an extremum estimators if it maximizes an objective function $Q_n(\theta)$, where Q_n may depend on a sample $(y_i)_{i=1}^n$ and a parameter θ.*

Note that the function $Q_n(\theta)$ is a random variable because it depends on the vector of observables. For the purpose of establishing asymptotic properties of an estimator, suppose that Q_n converges to a non-random $Q_0(\theta)$ as n becomes large and that the parameter θ_0 can be shown to maximize $Q_0(.)$. The next theorem is a fundamental result to show the consistency of an extremum estimator.

Theorem 2 *Suppose that (i) Θ is a compact subset of $\mathbb{R}^m$, (ii) $Q_n(.)$ is continuous in θ and (iii) $Q_n(.)$ converges uniformly to $Q_0(.)$, where θ_0 is the unique maximum of $Q_0(\theta)$. Then, $\hat{\theta}$ converges in probability to θ_0.*

Theorem 1 implies that, from the existence of the function $Q_0(.)$, $\theta = \theta_0$ is identified. Under some additional regularity conditions, an extremum estimator is asymptotically normal. A (heuristic) method of proof to show this point is useful because it will help understand how to explicitly calculate asymptotic variances in many models.

Suppose that θ_0 is interior in Θ and Q_n is twice-differentiable, so that the estimator will satisfy the first-order condition

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = 0.$$

Applying the mean-value theorem to the above term implies that the following expansion:

$$\underbrace{\frac{\partial Q_n(\hat{\theta})}{\partial \theta}}_{=0} = \frac{\partial Q_n(\theta_0)}{\partial \theta} + (\theta_0 - \hat{\theta})\frac{\partial^2 Q_n(\overline{\theta})}{\partial \theta \partial \theta'}, \tag{3.4}$$

where $\overline{\theta}$ is a value in-between θ_0 and $\hat{\theta}$. Reorganizing this equation and expanding by $\sqrt{n}$,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -(\underbrace{\frac{\partial^2 Q_n(\overline{\theta})}{\partial \theta \partial \theta'}}_{=B_n})^{-1} \underbrace{\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}}_{=A}. \tag{3.5}$$

Suppose that the hessian term B_n converges to a non-singular negative definite matrix

B. Because both $\hat{\theta}$ and $\overline{\theta}$ converge to θ_0, it is possible to estimate this matrix using the Hessian evaluated at $\hat{\theta}$ instead of $\overline{\theta}$.

The remaining term $\frac{\partial Q_n(\theta_0)}{\partial \theta}$ is expected to converge to zero as n becomes large but, once suitably expanded, will in many applications satisfy a central limit theorem and converge to a normal $N(0, \Sigma)$. For brevity, we do not provide here all technical conditions for asymptotic normality given that they are commonly satisfied in most applications; see Hayashi (2000) for details.

Theorem 3 *Under suitable regularity conditions, the extremum estimator is asymptotically normal with*

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, B^{-1}\Sigma B^{-1}). \tag{3.6}$$

3.3 M estimators and Maximum Likelihood

The challenge of general extremum estimators is that the derivation of Σ may not be straightforward. Fortunately, a special case of extremum estimators greatly facilitates this exercise and applies to estimation procedures such as maximum likelihood estimation (MLE).

Definition 5 *M estimators are extremum estimators such that the objective is to maximize a sample mean*

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} q(y_i; \theta). \tag{3.7}$$

Under MLE, denoting $f(.; \theta)$ as the p.d.f. of y, one maximizes:

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\theta), \tag{3.8}$$

so that the $q(.)$ function is the log likelihood of the observation y_i.

Estimating the hessian B is usually not a problem even for general extremum estimators because it can be recovered by taking the sample analogue $\hat{B}$, by calculating the

hessian for each observation and averaging over all observations. The remaining term $\frac{\partial Q_n(\theta)}{\partial \theta}$ is now an average, so that applying the central limit theorem

$$\sqrt{n}\frac{\partial Q_n(\hat{\theta})}{\partial \theta} \rightarrow N(0, \underbrace{Var(\frac{\partial q(\tilde{y};\theta_0))}{\partial \theta}}_{=\Sigma}). \tag{3.9}$$

This variance term Σ can then be estimated by taking the covariance associated to the sample analogue vector:

$$V = \begin{pmatrix} \frac{\partial q(y_1,\hat{\theta})}{\partial \theta} \\ \vdots \\ \frac{\partial q(y_n,\hat{\theta})}{\partial \theta} \end{pmatrix}. \tag{3.10}$$

This method is extremely important for applied analysis, because the researcher needs only to lay out a set of functions as a vector (or matrix if θ is multidimensional) and compute a covariance.

In the case of maximum likelihood, it is well-known that $B = \Sigma^{-1}$ so that, simplifying (3.6),

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \rightarrow_d N(0, B^{-1}). \tag{3.11}$$

The Hessian $B = -\mathbb{E}(\frac{\partial^2 log(\tilde{y},\theta_0)}{\partial \theta^2})$ is Fisher's information matrix and can be estimated with its sample analogue $\hat{B}$ by taking the second derivative of the log likelihood at each observation evaluated at $\hat{\theta}$. Over a wide class of problems, B^{-1} is the Cramér-Rao lower bound such that no other estimator can have lower asymptotic variance. Because the first-order condition on the log-likelihood is also a moment condition, MLE can be interpreted as an (asymptotically) optimal choice of moment conditions.

Maximum likelihood estimation admits a test statistic for nested models known as the likelihood ratio test. The objective of this test is to assess whether a simplified model, for example a model with fewer mechanisms, can explain the data. Let a restricted parameter space be denoted Θ_r with maximum likelihood estimate $\hat{\theta}_r$. Under the Null hypothesis

that $\theta_0 \in \Theta_r$, the test statistic

$$\lambda_{LR} = 2(Q_n(\hat{\theta}) - Q_n(\hat{\theta}_r)) \tag{3.12}$$

is asymptotically $\chi^2(d)$ with degree of freedom equal to the difference d in the dimensions of Θ and Θ_r.

Not all inference problems may require the researcher to compare a model against a more general one because even if the larger models fits data better, it may be significantly less parsimonious or cumbersome. In some cases, it may be of interest to compare two non-nested simple models to decide which of these two models is least misspecified.

A test statistic to answer this question is the Vuong (1989) test. Let us write the log likelihood ratio in (3.12) more generally as

$$LR_n = Q_n^1(\hat{\theta}^1) - Q_n^2(\hat{\theta}^2), \tag{3.13}$$

where Q_n^i (resp., $\hat{\theta}^i$) refers is the average log likelihood (resp., MLE estimate) of model i. Under the Null hypothesis that two competing models are equally far from the truth, in the sense of their expected log-likelihood ratio is zero, the test statistic

$$\lambda_v = \frac{LR_n}{\hat{w}_n\sqrt{n}} \tag{3.14}$$

is asymptotically standard-normal, where $\hat{w}_n$ is an estimate of the standard error of LR_n and can be estimated as in (3.10) by writing a vector of all likelihood ratios by observation, and calculating the standard-error of this vector.[10]

[10] Several adjustments to the numerator of the test statistic exist to improve the finite-sample properties of this test see, e.g., Vuong (1989). In particular, a common adjustment to the test statistic is to substract from the numerator $\frac{k_1-k_2}{2}\ln n$, where k_i is the dimension of the parameter space in model i.

3.4 Generalized Method of Moments

A second example of extremum estimators is the generalized method of moments (or GMM). Under GMM, the estimation procedure is based on M moment conditions with a $M \times 1$ vector function $g(.)$ that must satisfy:

$$\mathbb{E}(g(\tilde{y}, \theta_0)) = 0, \tag{3.15}$$

at the true value of θ. This suggests an objective function

$$Q_n(\theta) = (\frac{\sum g(y_i, \theta)}{n})' W_n \frac{\sum g(y_i, \theta)}{n}, \tag{3.16}$$

where W_n is an arbitrary positive definite weight matrix that assigns weights to the moments.

If W_n converges to a definite positive matrix, this estimator will typically satisfy the conditions for consistency and asymptotic normality. However, a good choice of weights should select moments that contain information about the parameters. Intuitively, moments that are more noisy should be given lesser weight and it can be shown that the optimal weight matrix, in the sense of obtaining the lowest asymptotic variance $B^{-1}\Sigma B$ is obtained with $W^{-1} = Var(g(\tilde{y}, \theta))$. This important result implies that, if the function $g(.)$ is known, the process of estimating optimal weights is straightforward by choosing

$$W_n = \frac{1}{n} \sum g(y_i, \theta) g(y_i, \theta)', \tag{3.17}$$

which can be computed by taking the covariance of a stacked matrix of moments:

$$\begin{pmatrix} g^1(y_1, \theta) & \dots & g^M(y_i, \theta) \\ \vdots & & \vdots \\ g^1(y_n, \theta) & \dots & g^n(y_n, \theta) \end{pmatrix} \tag{3.18}$$

This is similar to the method applied in (3.10) and involves recovering a matrix by evaluating the random variable at each observation. For applications in which $g(.)$ is additively separable in θ, the covariance matrix does not depend on θ and the covariance can be estimated without knowledge of θ. If, on the other hand, W_n depends on θ, all asymptotic properties of generalized method of moments can be obtained with a two-step procedure in which a first estimate of $\hat{\theta}$ is obtained using an inefficient weight matrix (i.e., the identity matrix) and then plugged in to obtain an estimate of the efficient weight matrix. Better *finite sample* properties can sometimes be achieved by repeating this procedure with new estimates to obtain more precise weight matrices; however, asymptotic properties of the estimation are not improved over a two-step procedure.

In what follows, assume that $W^{-1} = Var(g(\tilde{y}, \theta))$ is set to the optimal weighting matrix. Applying the asymptotic normality in (3.6) and simplifying with the optimal weight matrix,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, (G'WG)^{-1}), \tag{3.19}$$

where $G = \mathbb{E}(\frac{\partial g(\tilde{y}, \theta_0)}{\partial \theta})$ is the gradient of the moments and can be estimated using a sample analogue

$$\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g(y_i, \hat{\theta})}{\partial \theta}.$$

Generalized method of moments also admit a specification test known as the Hansen's J-test. Under the Null hypothesis that all moments are satisfied by the data, the J test statistic

$$J = nQ_n(\hat{\theta}) \tag{3.20}$$

is asymptotically $\chi^2(d)$ where d is equal to the difference between the number of moments and the number of parameters.

Some word of caution is required when interpreting a J test. Strictly speaking, rejection according to the J test means that the model is not a complete statistical explanation

of the data generating process as assessed by the moments. However, a J test rejection does not mean that the model has no useful implications or that the statistical differences between model and data invalidate all policy implications. Furthermore, responding to violations of the J test with a systematic search for a model that is not rejected, or by choosing moments that were found to fit well, invalidates the asymptotic distribution of the J test due to multiple hypothesis testing. Hence, not all researchers choose to compute this test statistic and failures of a J test are to be expected when a model is used to simplify a complex reality, what one might reasonably characterize as an intended objective rather than a weakness. In practice, a direct comparison of model-implied versus data moments is often more informative than a pass-fail test result in order to diagnose what specific data features are not fitted well by the model and guide future research.

3.5 Simulated Method of Moments

Many models do not have a closed-form solution; in these cases, writing an exact moment condition $\mathbb{E}(g(\tilde{y}, \theta))$ is infeasible and the researcher would instead use an approximation using a numerical solution of the model. For example, in the introductory example 1.2, the moments were approximated by simulating a dataset from a solution of the model: many dynamic model estimate with a full solution method would require simulation to recover the moment conditions. This method is known as simulated method of moments (SMM) and requires only minor adjustments to GMM. In principle, if the precision of the simulation becomes large as sample size increases, all asymptotics of GMM will apply. However, to improve the finite-sample properties of the estimation, it is relatively effortless to incorporate simulation noise.

In what follows, we focus on the most common application in which, for a dataset (y_i), the moment condition $\mathbb{E}(g(\tilde{y}) - \psi(\theta))$ is such that the functional form of $\psi(\theta)$ is unknown. Define $s \in [1, \ldots, S]$ simulated datasets $(y_i^s(\theta))$, obtained from solving the model numerically at θ and generating S simulated samples (or "fake" data). Computers use pseudo

random generators in which, by choosing a seed, the same simulation can be performed. Practically, each simulated sample is drawn with a different seed (or the simulation would be repeated) but the seed of each simulation should be held fixed across parameters, so that differences in simulation do not confuse the search for optimal parameters.[11]

Under SMM, the researcher finds a moment estimator $\hat{\theta}$ by solving

$$\hat{\theta} = argmin_{\theta} Q_n(\theta) \equiv g_n(\theta)' W_n g_n(\theta), \tag{3.21}$$

using a simulation for the unknown model prediction:

$$g_n(\theta) = \underbrace{\frac{1}{n}\sum_i h(x_i)}_{\text{Sample moment}} - \underbrace{\frac{1}{nS}\sum_{i,s} h(y_i^s(\beta)))}_{\text{simulated moment}}. \tag{3.22}$$

Because simulation noise in $(\tilde{y}_i^s(\theta))$ is independent from empirical sampling noise in (y_i), SMM can be understood as adding an exogenous noise term to the moment condition. Carrying over this term (Gourieroux, Monfort and Renault 1993), it can be shown that, holding S fixed in sample size, the asymptotic variance of the SMM estimator is

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d N(0, (1 + \frac{1}{S})(G'WG)^{-1}). \tag{3.23}$$

The J-test statistic must be adjusted for the fact that the objective function is now estimated with noise:

$$J = (1 + \frac{1}{S})^{-1} n Q_n(\beta). \tag{3.24}$$

The key benefit of SMM is that it opens the estimation of any model that can be solved numerically, i.e., without knowledge of analytical methods to compute theoretical properties. Modern computing has also made it much easier to solve larger models. On the

[11] If the seed is not held fixed, the search algorithm will automatically re-sample simulated datasets to optimize over the seed that best matches model features, overfitting the model by selecting an ideal random draw.

other hand, the method is several orders of magnitude more computationally intensive than other estimation procedures because it requires to solve and simulate the model for each parameter value, making it more challenging to estimate the model for a very large parameter space. For this reason, most applications are such that some parameters, identified without knowledge of the entire model, are estimated via a different method and plugged in the estimation.

3.6 Analytical Methods to Estimate Standard Errors

Up to this point, we assumed that there exists a function $g(.)$ characterizing a moment in closed-form; for any problem where this is the case, the methods developed above offer all that is needed to estimate parameters and derive their standard-errors. Below, we consider additional tools useful when the extremum estimator cannot be written as a straightforward moment condition or because the moment condition does not have a closed-form.

3.6.1 Delta Method

The delta method is a tool that can be used to derive the asymptotic properties of transformations of estimators satisfying asymptotic normality. It can be used in any application where a quantity of interest, for example, price efficiency or welfare, is a continuous function of the estimated parameters. Suppose that $\hat{\theta}$ is an estimator of θ satisfying a normal asymptotic expansion, that is,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Sigma). \tag{3.25}$$

The delta method states that for a differentiable function $h(.)$ with finite non-zero gradient, the estimator $h(\hat{\theta})$ is also asymptotically normal with:

$$\sqrt{n}(h(\hat{\theta}) - h(\theta_0)) \rightarrow_d N(0, (\frac{\partial h(\theta_0)}{\partial \theta})' \Sigma \frac{\partial h(\theta_0)}{\partial \theta}). \tag{3.26}$$

3.6.2 Clustered standard errors

The methods assume that observations are independent, which is rarely true in applications. For most applications, the most important form of correlation is within a group of observations (for example, within firm). A dataset has clusters is the dataset can be divided into separate subsamples $X_j = (y_{i,j})_{i=1}^m$,, where observations are independent across two clusters $j \neq j'$ but may be correlated within a cluster.

In the presence of clusters, it is sometimes possible to return to the baseline model by redefining an "observation" as a cluster X_j. The only difference is that the number of observations n now represents the number of separate clusters; for example, in maximum likelihood, $g(.)$ will be defined in terms of the log likelihood of the entire cluster (say, the time-series of a firm in the sample). Similarly, sampling can be adapted to sampling by cluster, or block bootstrap, so that, for example, the researcher draws firms instead of drawing individual observations.

Using this analogy, for generalized and simulated method of moments, the weight matrix can be adapted to clusters after reinterpreting assignment into a cluster as a random variable. Because observations are independent across clusters, the variance of the moment can be written as the sum of the moment variance in each cluster:

$$W^{-1} = \frac{1}{n} \sum_{j=1}^{J} Var(g_k(\theta, \tilde{y}), \tag{3.27}$$

where $Var(g_k(\theta, \tilde{y})$ can be estimated as the variance of the moment condition in cluster k by, as usual, stacking the moments of each observation in cluster k.

There is no simple analogue to these methods with two-way clusters, for example, if observations are correlated within firm and within year. If these correlations are important for the research question, a partial solution is to control for time heterogeneity by pre-cleaning variables with a fixed effect regression, detrending variables, or adding time fixed effects to the model. Alternatively, for known correlation structures, a demanding alternative is to simulate errors from the estimated data-generating process, i.e., using the statistical model to draw new error terms rather than the data. Unfortunately, this procedure may be inappropriate for stylized structural model whose purpose is not to fit the data in a statistical sense.

3.6.3 Influence Functions

Influence functions generalize the idea of deriving variances by stacking functions in (3.4). To form intuition about influence functions, consider the method used earlier by stacking $g(y_i, \theta)$ to derive the weight matrix: we refer to $g(y, \theta)$ the influence function of the moment and influence can be calculated more generally to derive standard errors of other estimators.

Consider an M estimator. Then, (3.5) can be expressed as

$$\hat{\theta} - \theta_0 \approx \frac{1}{n} \sum \underbrace{-\mathbb{E}(\frac{\partial^2 q(\tilde{y}, \theta_0)}{\partial \theta \partial \theta'})^{-1} \frac{\partial q(y_i, \theta_0)}{\partial \theta}}_{IF(y_i)}, \tag{3.28}$$

where the influence function $IF(y)$ can be stacked to obtain the asymptotic variance of the estimator. For example, in the case of maximum likelihood,

$$IF(y) = -B^{-1} \frac{\partial \ln f(y, \theta_0)}{\partial \theta}. \tag{3.29}$$

For the case of GMM estimators defined by moment conditions, using the mean-value

expansion of (3.5) yields:

$$\hat{\theta} - \theta_0 = -\underbrace{(G'WG)^{-1}}_{=B_n}\underbrace{GW\frac{1}{n}\sum g(y_i,\overline{\theta})}_{=A} \approx \frac{1}{n}\sum \underbrace{-(G'WG)^{-1}GWg(y_i,\theta_0)}_{=IF(y_i)}, \quad (3.30)$$

with $IF(y)$ defined as the influence function of the GMM estimator. Hence, a numerically equivalent method to estimate the asymptotic variance of the GMM estimator is to take the covariance of the pointwise empirical influence functions

$$\begin{pmatrix} \hat{IF}(y_1) \\ \dots \\ \hat{IF}(y_n) \end{pmatrix} = \begin{pmatrix} -(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}\hat{W}g(y_1,\hat{\theta}) \\ \vdots \\ -(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}\hat{W}g(y_n,\hat{\theta}) \end{pmatrix}. \quad (3.31)$$

The empirical influence function is $N \times M$, with dimension to the number of observations N times the number of moments M, so we can think about each column as the influence function of each moment.

The observation holds more generally to compute the asymptotic variance of a vector of estimators as long as their individual influence function is known. If IF_i denote the influence function of $\hat{\theta}_i$, the influence function of $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_M)$ is $IF(y) = (IF_1(y), \dots, IF_M(y))$ so that an asymptotic covariance can be estimated by stacking the empirical influence functions whereby each column i is a column vector $(\hat{IF}_i(y_j))_{j=1}^n$. For example, using this method, one can easily compute the asymptotic covariance of an estimator $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ such that $\hat{\theta}_1$ has been estimated by MLE while $\hat{\theta}_2$ has been estimated by GMM.

The analogue to the delta method for influence functions is the chain rule. Let $T(\theta_1, \dots, \theta_n)$ be a smooth function and such that $\hat{\eta} = T(\hat{\theta}_1, \dots, \hat{\theta}_M)$ is a function of M estimators $\hat{\theta}_i$ whose influence function IF_i are known. Then, the influence function of $\hat{\theta}$ is given by

$$IF = \sum \frac{\partial T(\hat{\theta}_1, \dots, \hat{\theta}_M)}{\partial \theta_i} IF_i(y). \quad (3.32)$$

Lastly, a common use of influence function is for carrying standard errors in multi-stage estimation. Assume that a moment $g(y, \theta; \delta)$ is a function of an estimator δ which can be estimated without knowledge of θ and is plugged in the moment condition. If the influence function of this estimator IF_δ is known, the influence function of θ is

$$IF(y) = IF_\theta(y) - \mathbb{E}(\frac{\partial g(\tilde{y}, \theta, \delta)}{\partial \delta}) IF_\delta(y), \tag{3.33}$$

where IF_θ is the influence function of θ if δ is known.

The examples below illustrate common estimators for which influence functions are easy to derive and known. To facilitate the exposition, it will be helpful to simplify (3.30) for the case of a single moment:

$$IF(y) = -\underbrace{\mathbb{E}(\frac{\partial g(\tilde{y}, \theta_0)}{\partial \theta})^{-1}}_{=G^{-1}} g(y, \theta_0). \tag{3.34}$$

Example 1: $g(y, m_\alpha) = y^\alpha - m_\alpha$ is the moment condition of an uncentered moment, then $G = -1$ in (3.34) and $IF(y) = y^\alpha - m_\alpha$.

Example 2: The variance σ^2 satisfies the moment condition:

$$g(y, \sigma^2; m_1) = (y - m_1)^2 - \sigma^2, \tag{3.35}$$

where m_1 can be plugged-in as a first-stage estimate. The influence function is then obtained from (3.33) as

$$IF(y) = (y - m_1)^2 - \sigma^2 + \underbrace{2\mathbb{E}(\tilde{y} - m_1)}_{\partial g / \partial m_1} \underbrace{(y - m_1)}_{IF_{m_1}} = (y - m_1)^2 - \sigma^2. \tag{3.36}$$

Example 3: The covariance $\mathcal{C}$ of $(\tilde{y}, \tilde{z})$ is given by a moment condition $g((y, z), \theta) = (y - m_y)(z - m_z) - \mathcal{C}$, where m_y and m_z are the means of each random variable. Using

the plug-in method,

$$IF(y,z) \;=\; (y-m_y)(z-m_z) - \mathcal{C} - \underbrace{\mathbb{E}(\tilde{z}-m_z)(y-m_y) - \mathbb{E}(\tilde{y}-m_y)(z-m_z)}_{=0}.$$

Example 4: Consider a linear model with $\tilde{y} = \tilde{x}\beta + \tilde{\epsilon}$. A least-squares regression can be stated as a moment condition $g((y,x),\beta) = x(y - x\beta)$ and has an influence function (Kahn 2015) given by:

$$IF(y,x) = \mathbb{E}(\tilde{x}'\tilde{x})^{-1}x(y - x\beta). \tag{3.37}$$

As a special case, the influence function of a conditional expectation can be obtained using (3.37) with x defined as an indicator variable for the conditioning event, and an auto-correlation coefficients by using x as lags of y.

3.7 Best Practices in Structural Models

As the preceding sections primarily emphasize various tools to estimate structural models, we relegate to this section some prescriptive notes relating to best practices when building and estimating a structural mode. These are intended as general principles rather than rules, and so should be taken as "Should" versus "Shouldn't" given that there may be specifics of a problem that require a particular approach or make some of these recommendations infeasible.

Items below refer to "Should" whenever possible.

1. Use a global search algorithm with bounded parameter space for baseline estimates; extend the bounds if the estimated parameter is on a boundary.

2. When using grids, check simulated data if the grid is sufficiently precise and policies in the simulation remain in the interior of the grid.

3. Check numerically for identification by simulating data from known parameters

near the baseline estimate and running the *same* estimation procedure as the baseline.

4. Use an optimal weight matrix when there are more moments than parameters but always use a method with reasonable precision for the weight matrix coefficients: as is well-known in portfolio theory, a poorly estimated covariance matrix can perform worse that a rule-of-thumb equal-weight matrix.

5. Explain the choice of moments in generalized method of moments and provide a theory as to why moments may plausibly identify each parameter.

6. In simulated method of moments, compute the theoretical moments accurately by using more observations (5 to 10 times) than in the dataset.

7. Non-parametric bootstrap is generally a *consistent* estimation method for the moment covariance matrix and standard-error on estimates (generally, any asymptotically normal estimator) but it is not always the most efficient. Avoid parametric bootstrap, i.e., simulating from the model, unless none of the other methods is feasible.

8. Examine interesting theoretical properties of the model, or of a simplified model, before it is formally estimated; understand economic mechanisms.

9. Do not use maximum likelihood without a sufficient model of unobserved noise terms.

10. Do not plug-in other estimates into an estimator without carrying first-stage standard-errors, using influence functions or bootstrap.

11. Write models that are internally-consistent. Internal consistency is usually more important than descriptive realism.

12. Avoid treating as exogenous parameters that are likely to change in response to a key counter-factual in another parameter.

13. Take special precautions when computing numerical derivatives: for a given step, ensure that a change in the precision of the solution algorithm has small effect on the numerical derivative.

14. Don't throw away your model because it fails a specification test; don't systematically tweak a model until it passes a test; failure to find evidence that a well-accepted model explains data is a meaningful result.

15. Avoid estimating too many structural parameters unless the objective function is very smooth and concave.

16. In SMM or GMM, don't use moments whose connection to parameters of interests is unclear.

17. Don't use a particular econometric method because it is used by other papers if there is another consistent estimator that is more suitable to the model.

18. Avoid long loops on an interpreted language like Matlab or R (use C or Fortran if using loop); think about vectorizing code.

19. Avoid motivating a paper as "writing a model about phenomenon X"; motivate it as answering a concrete question.

20. Don't use complicated econometrics or numerical methods for a small gain in efficiency unimportant for the research question, if there is a simpler method that can be replicated more easily.

4 Dynamic Models

4.1 Value Functions and Dynamic Programming

This section reviews numerical methods to estimate models featuring dynamic choices. The "full solution" approach, used in the introductory example in Section 1.2, involves solving for the optimal choice for any parameter value and match the theoretical predictions of the model to features of the data. The most common method to solve dynamic models is the principle of dynamic programming by Bellman (1966).

An economic agent operates over a time horizon $t = 0, \ldots, \infty$. The agent solves the following dynamic program:

$$V^*(x_0) = \max_{(d_t)} \mathbb{E}_0(\sum_{t=0}^{T} \beta^t u(x_t, d_t)|x_0), \tag{4.1}$$

where $x_t \in X$ is a bounded vector-valued state evolving according to a Markov process $Q(x'|x, d) = Pr(x_{t+1} \leq x'|x_t = z, d_t = d)$, d_t is a decision made each period which can be a function of past states $(x_t)_{t \leq k}$, and the payoff function $u(.)$ is continuous and bounded. Unfortunately, numerically solving this problem is difficult because it involves solving an objective in an infinite number of policies (d_t).

Equation (4.1) implies that, over all states x attained at some point of the process, the Bellman equation associated to this problem is

$$V^*(x) = \max_{d} u(x, d) + \beta \mathbb{E}(V^*(x')|x, d) \tag{4.2}$$

where, by convention, x' refers to the next period state and is drawn using $Q(.|x, c)$. The intuition for this reformulation of the problem is that the discounted value obtained in state x can be decomposed as making an optimal decision in one period plus the expected value of the state attained in the next period. This problem is simpler to solve because

the researcher need only solve for $V^*(.)$, a function with the dimensionality of the state space, and then can derive a policy function $d(.)$ by solving equation (4.2).

It is convenient to state (4.2) as a functional fixed point: the value function V^* must satisfy $\Gamma(V^*) = V^*$ where $\Gamma(.)$ is an operator that such that, for any function V, $\Gamma(V)$ is the function defined by:

$$\Gamma(V)(x) = \max_d u(x,d) + \beta \mathbb{E}(V(x')|x). \tag{4.3}$$

Although the value function solving (4.2) must be a fixed point, not all fixed points need to be an optimal solution. Therefore, a common step to guarantee the validity of the method is to ensure that $\Gamma(V) = V$ has a unique solution, which must then necessarily coincide with the solution V^* of the original problem.

A key result to guarantee this property is the contraction mapping theorem. In what follows, let $(\mathcal{M}, d)$ be a complete metric space such that $V^* \in \mathcal{M}$.

Theorem 4 *Suppose $\Gamma : \mathcal{M} \to \mathcal{M}$ is such that there exists $k \in (0,1)$ with $d(\Gamma(V), \Gamma(V')) \leq kd(V,V')$ for any $V, V' \in \mathcal{M}$. Then, Γ admits a unique fixed point V^* in $\mathcal{M}$ and, for any $V_0 \in \mathcal{M}$, $\Gamma^n(V_0)$ converges to V^*.*

Under the conditions of Theorem 4, Γ is said to be a contraction. For many dynamic programming problems, there exists a set of sufficient conditions that are usually easy to verify and which guarantee that the problem is a contraction.

Theorem 5 (Blackwell sufficient conditions) *Let $\Gamma : \mathcal{M} \to \mathcal{M}$ be defined on a set of bounded functions and such that:*

1. *(monotonicity) for $V_1, V_2 \in \mathcal{M}$ such that $V_1(x) \leq V_2(x)$, $\Gamma(V)(x) \leq \Gamma(V')(x)$;*
2. *(discounting) there exists $\beta \in (0,1)$ such that $\Gamma(V+a)(x) \leq \Gamma(V)(x) + \beta a$.*

Then, Γ is a contraction.

Blackwell's sufficient conditions hold generally for any dynamic programming problem. In the special case of Γ as defined in (4.3), Blackwell's conditions 1 and 2 are satisfied, so proving that a program is a contraction requires only to show that there is a reasonable class of bounded functions such that $\Gamma : \mathcal{M} \to \mathcal{M}$.

The contraction mapping theorem is also a proof method to derive theoretical properties of the model. Suppose that we want to show that the value function satisfies a property $\mathcal{P}$; for example, it may be concave or monotonic. To show this claim, it is sufficient to prove that $\Gamma(V)$ satisfies $\mathcal{P}$ for any $V \in \mathcal{M}$ satisfying $\mathcal{P}$. Formally, as long as $\mathcal{M} \cap \{V : V \text{ satisfies } \mathcal{P}\}$ is a complete metric space, the contraction mapping will imply that V^* is also in this set and will satisfy the property.

4.2 Numerical Analysis and Discretization

When implementing the solution to a dynamic programming on a computer, there are many methods available which differ in terms of computational speed and accuracy. We develop here an introduction to typical steps used in any implementation of these methods. Note that these steps should be viewed as a general template for computation. All coding languages offer user-supplied code for various standard tasks such as numerical differentiation to compute gradients and Hessians, or fast numerical integration.

The simplest (and usually quickest) method to solve a dynamic program is to solve the fixed point $\Gamma(V^*) = V^*$ in one step, by approximating

$$V^*(x) = \mathcal{V}(x; (a_i^*)_{i=1}^n) \tag{4.4}$$

using a flexible approximating family of functions indexed by a finite set of parameters $(a_i)_{i=1}^n$. For example, polynomial approximations, such as Chebychev or Legendre polynomials, can in principle approximate any smooth function.

The researcher can then maximize an objective function

$$(a_i^*)_{i=1}^n \in \min_{(a_i)_{i=1}^n} \int |\Gamma(\mathcal{V})(x;(a_i)_{i=1}^n) - \mathcal{V}(x;(a_i)_{i=1}^n)| dF(x) \tag{4.5}$$

where $F(.)$ is a distribution with full support over the set of states and n is chosen so that the minimum is small. The ideal choice of $F(.)$ would be to choose the steady-state distribution over the grid X^c; however, this is computationally complicated because it requires to simulate this distribution (by simulating many paths $(x_t)_{t=0}^T$). For this reason, a simpler approach is usually to use a uniform distribution over a bounded set of states.

The next methods are based on the second part of (4) and rely on the fact that, for any starting guess V_0, $\Gamma^n(V_0)$ should converge to the correct value function. The following algorithm is known as value function iteration.

1. At the first step $i = 0$, start from a guess for V^*, for example $V_0(x) = 0$.

2. Using the current guess V_i, evaluate the left-hand side of (4.3) using V_i and, solving for the optimal policy d_i, calculate a new guess

$$V_{i+1}(x) = \Gamma(V_i)(x) = \max_d u(x,d) + \beta \mathbb{E}(V_i(x')|x). \tag{4.6}$$

3. Check if the change in the value function $d(V_i, V_{i+1}) < d_0$ is below a certain convergence threshold d_0. If this condition is satisfied, stop the algorithm; if it is not satisfied, increase i by 1 and restart step 2.

An alternative to value function iteration is policy function iteration, which features fewer calls to the expensive step of maximizing (4.3) and replaces by the much cheaper step of computing values by repeating applications of the optimal policy. One can improve the update at step 2 using:

2'. In step 2, replace the update in (4.6) with the following step: for the optimal policy

d_i, obtain a value function:

$$V_{i+1}(x) = \mathbb{E}_0(\sum_{t=0}^{T} \beta^t u(x_t, d_t)|x_0 = 0, d_i), \tag{4.7}$$

by applying the policy function d_i.

For models with stochastic states, $V_{i+1}(x)$ in (4.7) can be approximated with Monte Carlo methods, by drawing a path of (x_t) and applying the decision rule d_i. This method requires additional coding for this step, but is usually faster because the additional step 2' yields a value function consistent with the policy without requiring a maximization.

Because computers do not operate on continuous spaces, the first step is to select a finite grid for the state space. To illustrate these methods, suppose that each state is two-dimensional $x = (y, z)$ such that $y' = f(x, y, z)$ is an endogenous state that depends on past states and past decisions, while z is a random exogenous state with c.d.f. $Q_z(z'|z)$. These methods can be readily extended to random endogenous states and additional dimensions. For a state space $X = Y \times Z$, consider a grid $X^g = Y^g \times Z^g$ with a total number of grid points $n_x = n_y \times n_z$.[12] The maximization in (4.3) is now performed on every point of the grid, and the new value function is updated on the grid.

A complication may arise when working with grids because, for a given decision d, x' may not be on the grid. There are several solutions to this problem. The first solution is to constrain the maximization to a grid for d such that $x' \in X^c$. Because this is often inconvenient, an alternative if $x' = f(x, d, z)$ has an inverse $\phi(.)$ such that $x' = f(x, \phi(x'), z)$ is to rephrase the problem such that the next state x', constrained to the grid X^c is a choice variable. As an example, in a neo-classical investment model in 1.2 , the next period capital is $x' = (1 - \delta)x + I$ where the investment I is a choice. This model can be rephrased by maximizing in next-period capital stock x' and substituting an

[12] While in the general model, x_{t+1} may be a random function of x_t,d_t and z_t, this formulation is without loss of generality in problems where $x_{t+1} = \phi(x_t, d_t, z_t) + \epsilon_t$. In this case, one can redefine an auxiliary i.i.d. exogenous variable $z_{t,2} = \epsilon_t$ and write $x' = f(x, (z, z_2), d)$.

implied investment $I = x' - (1-\delta)x$ required to achieve x'.

Unfortunately, this approach may not be feasible when x' is random or its relationship with choices is non-trivial. In this case, an alternative is to use an interpolation methods where for any $x' \notin X^g$, the value $V(x')$ is obtained as an average of values on the grid. Various implementations of these algorithms exist for all programming languages. When interpolating, it should be verified that there is no x' outside of the range of the grid since, then, the interpolation is simply extrapolating using its own assumed functional form.

4.3 Vectorized Code

The methods developed so far involve a loop over all states. Loops are fast and efficient with uninterpreted languages, such as C or Fortran, but long loops are unsuitable for interpreted languages. Hence, when working with interpreted languages it is almost always preferable to vectorize code in order to avoid loops. When vectorizing, the maximization is conducted simultaneously over a matrix of all state-contingent payoffs instead of being maximized sequentially in a loop.

In vectorized form, the maximization

$$\max_d \underbrace{u(x,d) + \beta \mathbb{E}(V(x')|z)}_{=M(i,k)},$$

is now represented as a matrix whose first dimensions are the grid points of the state x_i and the next dimension is a grid point for the decision $d_k \in D^g$. The matrix is two-dimensional when the state and the decision are vectors but otherwise has dimensions equal to the sum of the dimension of the states and the dimension of the decisions. For any (i,k), the matrix should contain the payoff in brackets. Interpreted language are able to quickly maximize the matrix in its decision k to perform one iteration of the value function without loop. In some cases with large matrices, vectorized code can be further accelerated using a GPU.

To illustrate, consider the special case of this problem in which the state has two dimensions (y, z) and the decision $I = \phi(x')$. Then, the above maximization can be written as:

$$\max_k \underbrace{u(y(i), z(j), \phi(x(k))) + \beta \mathbb{E}(V(x(k), z')|z(j))}_{=M(i,j,k)}.$$

The current period payoff can be directly written by evaluating $u(y(i), z(j), \phi(x(k)))$. The second part, which is a conditional expectation, can be obtained by calculating $Q_z V^M(i, j)$ where $V^M(i, j) = V(x(i), y(j))$ is the value function in matrix form. The updated value function can then be calculated as

$$V^M(i, j) = \max_k M(i, j, k),$$

which can be performed on any interpreted language.

If a state space $x = (z, y)$ contains an exogenous state whose distribution $Q^z(z'|z)$ is an auto-regressive process $z_{t+1} = \rho z_t + (1 - \rho) z_{t-1} + \epsilon_t$, where ϵ_t is a distribution with finite moments, the Tauchen (1986) method yields well-known approximations for the state space $z \in Z_d$ and the transition probabilities $Q_z^d(z' \in Z_d|z \in Z_d)$. This method is commonly-used for cases in which ϵ_t is normal and is easily adapted to functions of normals (lognormal), but the approximation can be applied to more general distributions.[13]

There are also approximations that are less precise for finite grids but are computationally faster. These methods can be easily adapted to auto-regressive processes but are most intuitive in the i.i.d. case where z_t has an i.i.d. distribution $F(.)$. Given a grid of I quantiles $\Lambda_d = (F^{-1}(i/(I + 1)))_{i \in [1,I]}$, standard results guarantee that F can be approximated as a discrete uniform distribution with support on Λ_d. In models where z_t

[13]Most textbooks on numerical methods advise not to use approximation methods relying on normalized densities, given that they are known to be inferior to the Tauchen method (Judd 1998a). Such methods usually involve evaluating the density at a point of each discretized interval (for example, the mid-point) and normalizing by the sum of densities so that probabilities add to one. Unlike the Tauchen method, which is based on cumulative probability functions, these methods can become imprecise when densities are very small or vary quickly.

is the only random shock, it is possible to compute the expectation $\beta\mathbb{E}(V_i(x')|x)$ by first evaluating $V_i(x')|x)$ and taking an average over all values of z' on the grid.

The endogenous variable presents more significant challenges because, a-priori, a suitable grid is unknown and may vary as a function of parameter values. Most authors will manually improve the grid as an estimation converges by checking that, in simulations, states tend to be reached in areas interior to the grid and the grid is sufficiently precise so that further refinements would have small effects on the optimal policy.

It is sometimes possible to use a dynamic grid guided by solving a simpler problem that yields upper and lower bounds on the grid or populate a tighter grid in more influential parts of the state space. A notable challenge with dynamic grids is that, in simulated method of moments, they can interfere with the process of calculating gradients of moment conditions. A derivative may be affected by changes in the grid and cause the gradient to be either too high or too low; it is also difficult to compare the suitable precision of a dynamic grid with the suitable precision of a numerical gradient. For this reason, it is preferable to use dynamic grid initially to search for an estimate, but then revert to a precise fixed grid for the final local search and computations of gradients. Ideally, a numerical gradient should not vary when increasing the precision of the grid.

4.4 Application to Dynamic Conditional Choice Probabilities

4.4.1 Observed States

The standard dynamic discrete choice model was proposed by Hotz and Miller (1993) as an alternative to solving the dynamic model in problems where states and choices are observable by the researcher. Hotz and Miller observe that the value function, which is usually numerically solved under conventional full-solution approaches, can be inverted from the probability of each choice conditional on each observed state. By avoiding the computational step of solving the value function, the approach can accommodate richer

models than full-solution methods. It has received considerable attention in industrial organization and marketing, where the large set of individual and brand covariates can make it difficult to solve models. We develop a below a brief treatment of conditional choice probabilities (CCP), first when states are perfectly observed and, then, in its recent development with unobserved states (Arcidiacono and Miller 2011) provided that the state is not affected by choices.

A firm indexed by $n \in [1, N]$ operates over an infinite horizon $t = 0, \ldots, \infty$ and receives current period payoffs

$$\underbrace{a_j + b_j x_{nt} + c_j s_n + d_j x_{nt} s_n}_{\equiv u_j(z_{nt})} + \epsilon_{njt}, \tag{4.8}$$

where (i) $\theta = (a_j, b_j, c_j, d_j)_{j \in J}$ is a set of parameters to be estimated, (ii) each $j \in J$ indicates a choice, (iii) $z_{nt} \equiv (x_{nt}, s_n)$ is a state from a space $Z = X \times S$ with $n_z \equiv n_x \times n_s$ elements drawn according to a discrete distribution $f_j(z_{nt+1}|z_{nt})$, and (iv) the vector $\epsilon_{nt=(\epsilon_{n1t},\ldots,\epsilon_{nJt})}$ is period noise i.i.d. over time and follows a type I extreme Value distribution with c.d.f. [14]

$$G(\epsilon_{nt}) = \exp\left(-\sum_j \exp(-\epsilon_{njt})\right). \tag{4.9}$$

To set ideas, a simplified example of this model is from Rust (1987). Each period, the firm receives a payoff $u_0(x_{nt}) = \epsilon_{n0t}$ from replacing its bus, or can, alternatively, keep its current bus with age x_{nt} for a payoff $u_1(x_{nt}) = bx_{nt} + \epsilon_{n1t}$, where bx_{nt} is the expected cost of repairs; ϵ_{njt} reflects idiosyncratic factors of bus engines. Conditional on replacement, $j = 0$, the age of the bus is $x_{nt+1} = 1$ and conditional on keeping the bus, the age advances to $x_{nt+1} = x_{nt} + 1$.

[14] The type I extreme value assumption implies that the estimation reduces to a standard logit, simplifying the computation of the likelihood but is otherwise not to critical to these methods. With minor changes to the equations, the model can accommodate correlation within-period errors (i.e., replacing the logit with a nested logit), see, e.g., Arcidiacono and Miller (2011) equation (3.18).

The firm sets a Markov decision rule $d(z,\epsilon) \equiv (d_1(z,\epsilon), \ldots, \ldots d_J(z,\epsilon))$, with each $d_j \in \{0,1\}$ and $\sum_j d_j(z,\epsilon) = 1$, to maximize

$$V(z) \equiv \mathbb{E}\left(\sum_{t=0}^{\infty} \sum_{j=1}^{J} \beta^t d_j(z_{nt}, \epsilon_{njt})[u_j(z_t) + \epsilon_{jt}] | z_{n1} = z \right). \tag{4.10}$$

Standard results in the literature demonstrate that β is in general not identified (Rust 1987, Magnac and Thesmar 2002), so we shall assume throughout that $\beta \in (0,1)$ is known. For any z and j, the conditional choice probability (CCP) is

$$p_j(z) = \int d_j(z,\epsilon) dG(\epsilon),$$

with associated vector $p(z) = (p_1(z), \ldots, p_J(z))$. Let $v_j(z)$ denote the lifetime payoff from action j *without the period noise*:

$$v_j(z) \equiv u_j(z) + \beta \sum_{z'} V(z') f_j(z'|z). \tag{4.11}$$

Because $p_j(z) = Pr(j \in \text{argmax}_{j'} v_{j'}(z) + \epsilon_{j'})$, standard extreme value theory implies (McFadden 1973) that

$$p_j(z) = \frac{e^{v_j(z) - v_1(z)}}{1 + \sum_{j'} e^{v_{j'}(z) - v_1(z)}}, \tag{4.12}$$

which implies the structure of a multinomial logit in which the probability of choice j is written of the value function for choice j net of base choice with index $j = 1$. However, because the mapping between the matrix $v_j(z) - v_1(z)$ and the period utility parameters in (4.8) is unknown, this step does not yield yet an estimation procedure for the parameters of interest. Making this mapping explicit is the focus of the next steps.

When (4.9) holds, Hotz and Miller (1993) show that

$$V(z) - v_j(z) = \gamma - \ln(p_j(z)), \tag{4.13}$$

where $\gamma \approx 1.781$ is Euler's constant. Reinjecting the above equation evaluated at $j = 1$ into (4.11),

$$v_j(z) = u_j(z) + \beta \sum_{z'} f_j(z'|z)(v_1(z') + \gamma - \ln(p_1(z'))), \tag{4.14}$$

which implies that all $v_j(z)$ can be written as a function of a benchmark choice $v_1(z)$. As in static discrete choice, utilities are identified up to a base choice. Hereafter, we normalize $u_1(z) = 0$, so that $u_j(z)$ is the incremental utility of choice j over base choice 1 conditional on state z.

Evaluating (4.14) at $j = 1$ implies the following system of equations for $v_1(z)$:

$$v_1(z) = \beta \sum_{z'} f_1(z'|z)(v_1(z') + \gamma - \ln(p_1(z'))). \tag{4.15}$$

If the state z_{nt} is observed, the model can be readily estimated with a three-step procedure making use of the economic restrictions in (4.12)-(4.15). We refer to (d_{njt}, z_{njt}) as the empirical sample of decisions and states.

Step 1. Obtain first-stage estimates from the transition probabilities $f(z'|z)$ and the CCP $p_1(z')$. For a small number of states in which each state is visited sufficiently often, one can use a bin estimator:

$$\hat{f}_j(z'|z) = \frac{\sum_{t,n} 1_{z_{nt}=z, z_{nt+1}=z', d_{njt}=1}}{\sum_{t,n} 1_{z_{nt}=z, d_{njt}=1}} \tag{4.16}$$

$$\hat{p}_j(z) = \frac{\sum_{t,n} 1_{z_{nt}=z, d_{njt}=1}}{\sum_{t,n} 1_{z_{nt}=z}}. \tag{4.17}$$

For a larger number of states in which some bins may not be precisely estimated, one may alternatively use a parametric form, e.g., $\hat{f}_j(z'|z) = \phi(\beta_0^j + \beta_1^j z + \beta_2^j z')$ and $\hat{p}_j(z') = \phi(\nu_0^j + \nu_1^j z + \nu_2^j z')$, where $\phi(.)$ a known cumulative density function, and estimate

the parameters by maximizing the log likelihood:

$$\mathcal{L}_{\beta\nu} \equiv \underbrace{\sum_{t,n,j} d_{njt} \ln \hat{f}_j(z_{nt+1}|z_{nt})}_{\mathcal{L}_\beta} + \underbrace{\sum_{t,n,j} d_{njt} \ln \hat{p}_j(z_{nt+1}))}_{\mathcal{L}_\nu}. \tag{4.18}$$

Note that, in the right-hand side of (4.18), $\mathcal{L}_\beta$ and $\mathcal{L}_\nu$ can be estimated separately.

Step 2. Plug these first-stage estimates into (4.15) and solve for $v_1(z)$ from a linear system of n_z equations in n_z unknowns:

$$v_1(z) = \beta \sum_{z'} \hat{f}_1(z'|z)(v_1(z') + \gamma - \ln(\hat{p}_1(z'))). \tag{4.19}$$

Step 3. Having solved for $v_1(z)$ in step 2, compute the value from the other choices $v_j(z)$ from (4.14). The model can then be estimated by maximizing the likelihood associated to equation (4.12):

$$\mathcal{L} = \prod_n \underbrace{\prod_{t,j} \left(\frac{e^{v_j(z_{njt}) - v_1(z_{njt})}}{1 + \sum_{j'} e^{v_{j'}(z_{njt}) - v_1(z_{njt})}}\right)^{d_{njt}}}_{\equiv l_n(s_n)}. \tag{4.20}$$

4.4.2 Unobserved Heterogeneity

Suppose next that the state $s_n \in S$ is unobservable to the econometrician and s_n is drawn with probability $\pi(s_n)$. The intuition for adapting CCP to unobserved heterogeneity is that states imply different likelihoods over observed choices, so that the distribution of the unobserved states can be (in principle) estimated by maximizing a likelihood. However, the estimator of the CCP in (4.17) is no longer feasible (as it depends on s_n).

Arcidiacono and Miller (2011) propose an iterative method relying on the fact that (i) if the CCP *were* known, one could easily estimate the likelihood that a firm is in state s from its choices, the conditional probability of each choice (i.e., CCP) and Bayes rule

- also known as expectations maximization (EM) algorithm; and (ii) if the likelihood of each state is known for each firm, we can derive the CCP by weighting choices by the probability of the state.

To implement this method, let first $p^0(z)$ be an initial vector of conjectured CCP, a conjectured vector θ^0 and a conjectured probability of states $\pi^0(s)$. Given that s_n is constant within firm, the transition $f(z'|z)$ can still be painlessly estimated from (4.16) or a related method. The following steps are replacements to steps 1-3 in the baseline model without unobserved heterogeneity.

Step 1'. Solving the system of equations in step 2, the function $l_n(s)$ can be calculated for any state s and implies that the probability that firm n is in state s is

$$q_n^1(s) \equiv \frac{\pi^0(s) l_n(s)}{\sum_{s'} \pi^0(s') l_n(s')}, \tag{4.21}$$

and the distribution of states can be updated to $\pi^1(s) = \sum_n q_n^1(s)/N$.

Step 2'. Rewrite (4.17) by weighting the choices of each firm by the probability it is in state s:

$$p_j^1(z) = \frac{\sum_{t,n} q_n^1(s) 1_{x_{nt}=x, d_{njt}=1}}{\sum_{t,n} q_n^1(s) 1_{x_{nt}=x}}. \tag{4.22}$$

Step 3'. The likelihood can then be adapted by integrating over all states, that is,

$$\mathcal{L} = \prod_{n,s} q_n^1(s) l_n(s), \tag{4.23}$$

and maximized in θ to obtain an updated θ^1. Estimates for $\hat{p}(s)$, $\hat{q}_n(s)$, $\hat{\pi}(s)$ and $\hat{\theta}$ are then obtained by iterating over steps 1' to 3' until convergence.

4.5 Renewal Problems with an Application to Auditing

Discrete choice models have been recently applied to estimating models of audit choice (Gerakos and Syverson 2015, Gerakos and Syverson 2017, Guo, Koch and Zhu 2017, Cheynel and Zhou 2020b, Guo, Koch and Zhu 2021) given that preferences of clients for audits with various characteristics, such as auditor size, experience, specialization, can be approached as the purchase of a differentiated durable credence good (Causholli and Knechel 2012).

Consider a simplified version of the dynamic discrete choice model in Cheynel and Zhou (2020b), where client and auditors consider the current and future value of their existing relationship. In their model, a renewal occurs when the client changes auditor. Formally, a renewal problem is defined as transitions such that making the base choice $j = 1$ leads to a future state $\underline{z}$ that does not depend on the current state (for example, a replacement choice). Under this assumption, equation (4.14) simplifies to

$$v_j(z) = u_j(z) + \beta v_1(\underline{z}) + \beta\gamma - \beta \sum_{z'} f_j(z'|z) \ln(p_1(z')), \tag{4.24}$$

which implies that

$$v_j(z) - v_1(z) \quad = \quad u_j(z) - \beta \sum_{z'} f_j(z'|z)) \ln(p_1(z')) + \beta \ln p_1(\underline{z}). \tag{4.25}$$

When s is observable, this model can be estimated by running a multinomial logit on observations in state z on a constant, x, s and xs, to recover estimates $\hat{\eta}_j$, $\hat{b}_j$, $\hat{c}_j$ and $\hat{d}_j$, where the last three estimates represent the parameter of the preference b_j, c_j and d_j. The remaining preference parameter in (4.8) a_j can then be estimated as

$$\hat{a}_j = \hat{\eta}_j + \beta \sum_{z'} \hat{f}_j(z'|z)) \ln(\hat{p}_1(z')) - \beta \ln p_1(\underline{z}), \tag{4.26}$$

where the last terms in (4.26) capture the dynamic choice.

Suppose next s is not observable. Let the conjectured CCP and conjectured parameters of the period utility be denoted by $p_1^1(z)$ and θ^1, respectively, and such that $u_j^1(x,s) \equiv a_j^1 + b_j^1 x + c_j^1 s + d_j xs$ refers to the conjectured period utility. The likelihood $l_n(s)$ is written in closed-form as

$$l_n(s) = \prod_{t,j} \left(\frac{e^{u_j^1(x_{nt},s) - \beta \sum_{z'} \hat{f}_j(z'|z_{nt}) \ln(p_1^1(z')) + \beta \ln p_1^1(\underline{z})}}{1 + \sum_{j'} e^{u_{j'}^1(x_{nt},s) - \beta \sum_{z'} \hat{f}_{j'}(z'|z_{nt}) \ln(p_1^1(z')) + \beta \ln p_1^1(\underline{z})}} \right)^{d_{njt}}. \tag{4.27}$$

4.6 Application to Investment Theory

This section describes the use of a quantitative general equilibrium model to get an inference of different accounting systems for allocative efficiency. Unlike traditional productive efficiency as modelled in many investment models, where a firm raises a varying amount of capital at a fixed unit cost, allocative efficiency explicitly considers how capital is re-allocated between firms (Hwang and Kim 2019, Cheynel and Bertomeu 2020, Breuer 2021).[15]

The current introductory treatment follows the quantitative general equilibrium analysis in Choi (2021), which builds on a simplified version of the David, Hopenhayn and Venkateswaran (2016) general equilibrium model with heterogeneous firms under imperfect information and the accrual accounting systems in Nikolaev (2019). Using this approach, Choi (2021) studies the impact of accrual accounting systems, relative to cash accounting systems, on aggregate productivity and output in the economy.

In (David et al. 2016), the relation between imperfect information and resource misallocation under the equilibrium condition is given by

$$\frac{dy}{d\overline{V}} = -\frac{1}{2}\theta \frac{1}{1-\alpha.}, \tag{4.28}$$

where y represents an aggregate output, $\overline{V}$ represents a summary measure of imper-

[15] We gratefully thank Jungho Choi for his generous assistance on this Section.

fect information, θ represents the elasticity of substitution between labor and capital, and α represents capital share. The summary measure of imperfect information ($\overline{V}$) has a negative impact on the (ex-post) allocative efficiency of capital and labor because imperfect information might lead to (ex-post) high performing firms utilizing limited resources ex-ante and (ex-post) low performing firms utilizing excess resources ex-ante as well.

Choi (2021) estimates $\overline{V}$ from equation (4.29) below, which expresses uncertainty in terms of fundamental economic uncertainty, as well as various uncertainty components related to correlation, managerial information, cash flows and accruals:

$$\overline{V} = \underbrace{\rho^2(\frac{\sigma_s^2}{\sigma^2+\sigma_s^2})^2(\frac{1}{\sigma_{ae}^2}+\frac{1}{\sigma_{cf}^2}+\frac{\rho^2}{\sigma^2+\sigma_s^2}+\frac{1}{\overline{V}})^{-1}}_{\text{Uncertainty about historical productivity}} + \underbrace{(\frac{1}{\sigma^2}+\frac{1}{\sigma_s^2})^{-1}}_{\text{Uncertainty about current productivity}}. \tag{4.29}$$

The first term in the equation captures uncertainty about historical productivity", which acknowledges an imperfect performance measure and the role of accounting systems in the economy, and the second term is the uncertainty about current productivity which contains the innovation to productivity. These terms can be estimated from properties of cash flows and accruals.

5 Structural Models of Agency Theory

5.1 A Canonical Model

Agency theory describes situations in which a principal (e.g., the firm) hires an agent whose actions may affect the outcome of a project. These actions are unobservable or non-contractible so that, to elicit adequate actions, the principal must offer a compensation contract that conditions pay on a performance measure. This framework was popularized by Jensen and Meckling (1976) and, since then, has been a leading approach to

examine optimal contracts and the distortions due to asymmetric information in employment relationships.

While this work has been enormously influential in generating qualitative insights, the equations of the theory have had, to this date, very limited use in shaping the design of optimal contracts or quantitatively measuring agency distortions. The lack of research using the actual structural equations of these models poses two fundamental questions. First, from a strict perspective of philosophy of science, the theory cannot be simultaneously true when it makes qualitative predictions and false when its equations are applied to empirical samples. Second, the many versions of this framework combined with loose directional implications make it very difficult to test the theory, as it is easy to explain a qualitative feature of the date, or its opposite, using a some plausible specification of the agency problem. A structural model, by contrast, takes the entire model seriously which facilitates a coherent joint test of all the implications of the model, including the soundness of its quantitative implications.

We consider here the canonical model of Holmström (1979). Due to its simplicity and versatility, this model is the most commonly-used version of principal-agent model and we shall follow its implementation in a structural model in Margiotta and Miller (2000), Gayle and Miller (2009, 2015), Gayle, Golan and Miller (2015), and Gayle, Li and Miller (2018, 2021).

Assume that the agent is risk-averse and achieves a utility function $u(w;e)$, which depends on pay w and effort e with first derivatives $u' > 0$, $u'' < 0$ and $u(w;1) < u(w;0)$ captures a cost of effort. Suppose that the utility function satisfies the Inada conditions such that $w(x)$ is interior. The agent is offered a compensation schedule $w(.)$ and then privately chooses $a \in \{0,1\}$, where $e = 1$ indicates high effort and $a = 0$ indicates shirking. Then, the firm generates a contractible signal on output $\tilde{x}$ drawn from a distribution with density $f(x|a) > 0$ and such that the monotone likelihood ratio property (MLRP) holds, i.e., $f(x|0)/f(x|1)$ is decreasing in x.

The agent has an alternate occupation which generates a utility level $u(\underline{w}; 0)$ and, therefore, the contract must achieve a minimum utility level given in the participation constraint (PC) below:

$$\int f(x|1)u(w(x);1)dx \geq u(\underline{w};0). \tag{PC}$$

The contract must also induce an optimal choice $a = 1$, implying an incentive-compatibility condition (IC):

$$\int (f(x|1)u(w(x);1) - f(x|0)u(w(x);0))dx \geq 0. \tag{IC}$$

The principal is risk-neutral and solves the following program

$$\max_{w(.)} \int f(x|1)(x - w(x))dx$$

s.t. (PC) and (IC).

In what follows, assume that this program has an optimum policy. Specifically, Mirrlees (1999) shows that programs in which sufficiently low events indicate shirking, i.e, such that $f(x|0)/f(x|1)$ is unbounded from above, and the possible penalties imposed by the principal are unrestricted, i.e., $u(.)$ is unbounded from below, imply that there is a sequence of contracts that converges to the principal surplus without the (IC). These contracts are, in the limit, almost surely flat with a very low $w(.)$ for extreme events. This type of contracts does not appear to match observed contracts. As a special case, an optimal contract does not exist in the common context of exponential (CARA) utility functions with a normally-distributed performance measure $\tilde{x} \sim N(a, \sigma^2)$. To avoid this problem, most implementations of this framework assume that $f(x|0)/f(x|1)$ is bounded from below.

Denote λ the Lagrange multiplier associated to (PC) and μ the Lagrange multiplier

associated to (IC). The Lagrangian of the problem is

$$\mathcal{L} = \int (x - w(x)) f(x|1) dx - \lambda (u(\underline{w}; 0) - \int f(x|1) u(w(x); 1) dx)$$
$$-\mu (\int (f(x|0) u(w(x); 0) - f(x|1) u(w(x); 1)) dx.$$

Differentiating this Lagrangian with respect to $w(.)$,

$$\frac{\partial \mathcal{L}}{\partial w(x)} = -1 + \lambda u'(w(x); 1) + \mu (u'(w(x); 1) - \frac{f(x|0)}{f(x|1)} u'(w(x); 0)) = 0. \tag{5.1}$$

The first-best denotes the problem in which effort is chosen directly by the principal (or is contractible). Dropping (IC) and setting $\mu = 0$ in (5.1) implies that $\lambda u'(w(x); 1) = 1$ so that the compensation will be constant to avoid imposing unnecessary risks on the agent. In the second-best program, by contrast, a flat contract cannot be optimal and, therefore, the Lagrange multiplier $\mu > 0$ and the term $u'(w(x); 1) - \frac{f(x|0)}{f(x|1)} u'(w(x); 0)$ captures the incentive component. The two constraints (PC) and (IC) must be met at equality, so that (IC) can be equivalently written as

$$\int f(x|0) u(w(x); 0) dx = u(\underline{w}; 0). \tag{IC'}$$

We are now equipped with sufficient theory to identify and estimate the primitives of this model given observations from the performance measure and wage payments $(\tilde{x}, w(\tilde{x}))$. However, Gayle and Miller (2015) show that this model is not identified without placing additional structure on the model. To give intuition for their insight, solving the model yields an implication in the form of a function $w(x)$ but this function can only serve to identify the unknown function $f(x|0)$ from the first-order condition (5.1) as long as $u(w; a)$ and $\underline{w}$ are known. A more restrictive model specification, as discussed below, is required to estimate this model.

5.2 Semi-Parametric Identification with Additive Cost of Effort

Our interest will be to examine whether primitives of the model, namely, the utility function of the manager (inclusive of the effort cost) and the distribution of the performance measure can be recovered from an empirical sample containing observations of the performance measure and pay $(x_i, w_i)_{i=1}^n$. In this section, we identify the model under the following assumptions:

(A1) the cost of effort is additive, such that $u(w; a) = u_0(w) - ca$ with $c > 0$;

(A2) denoting $\tilde{x}_a$ as the performance measure conditional on effort a, $\tilde{x}_0 = \alpha_0 + \alpha_1 \tilde{x}_1$;

(A3) utility functions are in the HARA (hyperbolic absolute risk-aversion) class:

$$u_0(w) = \frac{1-\gamma}{\gamma}(\frac{\beta w}{1-\gamma} + \nu)^\gamma. \tag{5.2}$$

Under (A1), equation (5.1) simplifies to the following familiar characterization of an optimal contract in Holmström (1979).

$$\frac{1}{u_0'(w(x))} = \lambda + \mu(1 - \frac{f(x|0)}{f(x|1)}). \tag{5.3}$$

(A2) further implies that the likelihood ratio can be written in terms of the observable $f(x|1)$ and (α_0, α_1) where

$$f(x|0) = \frac{1}{\alpha_1} f(\frac{x-\alpha_0}{\alpha_1}|1) \tag{5.4}$$

Following (A3), utility functions in the HARA class have been used in agency models (Hemmer, Kim and Verrecchia 2000, Bertomeu 2014) and imply a marginal utility

$$u_0'(w) \quad = \quad \beta(\nu + \frac{\beta w}{1-\gamma})^{\gamma-1}. \tag{5.5}$$

Having characterized model restrictions (PC), (IC) and (5.6), there are several ap-

proaches available to estimate the parameter of the model.

Semi-parametric Approach. In practice, one can use a first-stage non-parametric kernel estimate of $\hat{f}(x|1)$, which is easily obtained from (x_i). Plugging (5.5) into (5.3) then implies the following characterization of the optimal wage:

$$w(x) = \frac{1-\gamma}{\beta}((\beta(\lambda + \mu(1 - \frac{1}{\alpha_1}\frac{f(\frac{x-\alpha_0}{\alpha_1}|1)}{f(x|1)}))^{1/(1-\gamma)} - \nu). \tag{5.6}$$

Replacing $f(x|1)$ by $\hat{f}(x|1)$, the parameters $\theta = (\gamma, \beta, \lambda, \mu, \alpha_0, \alpha_1, \nu)$ can be estimated by non-linear least squares, minimizing the squared difference between model and actual compensation:

$$Q_n = \frac{1}{n}\sum_{i=1}^{n}(w(x_i) - w_i)^2. \tag{5.7}$$

The two remaining parameters $(\underline{w}, c)$ can then be obtained by plugging the estimated wage $\hat{w}(x)$ from (5.6) into (PC) and (IC) and solve for the coefficient so that these equations are satisfied.

To avoid the loss of efficiency due to plugging in these estimates to recover $(\underline{w}, c)$, a single-step version of this approach is add the (PC) and (IC) as a matrix of stacked moments:

$$Q_n = G_n' W G_n, \tag{5.8}$$

where:

$$G_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(w(x_i) - w_i)^2 \\ \frac{1}{n}\sum_{i=1}^{n} u_0(w_i) - c - u_0(\underline{w}) \\ \frac{1}{n}\sum_{i=1}^{n} p(x_i)u_0(w(x_i)) - u_0(\underline{w}), \end{pmatrix} \tag{5.9}$$

where $p(x_i)$ indicates a discretization of the output conditional on shirking and can be computed by integrating $\hat{f}((x-\alpha_0)/\alpha_1|1)$ over a grid.[16] One can use the identity as a

[16] An alternative is to use $p(x_i) \equiv \hat{f}((x_i-\alpha_0)/\alpha_1|1)/\sum_{j=1}^{J}\hat{f}((x_j-\alpha_0)/\alpha_1|1)$; however, this type of

weight matrix or, for more precision, a weight matrix that is a function of the moments. The variance of the moments is a complicated function of first-step plug-in estimates $\hat{f}(.)$ and its analytical expression is non-trivial. The sampling methods in section 2.4 can nevertheless offer a feasible computational method to weight moments.

Parametric Approach. A feasible approach to estimate this model involves strengthening (A2) with a parametric assumption for $f(x|a)$. For example, Gayle and Miller (2009) assume that $f(x|a)$ is the p.d.f. of a Normal distribution $N(m_a, \sigma^2)$ and $u_0(w) = -e^{-rw}$ is exponential. The eight model parameters be given by $\theta = (m_1, \sigma, c, \underline{w}, \gamma, r)$ can then be estimated by numerically solving the model and matching model-implied moments and data moments as in Section 2. Here, possible moments include the mean and variance of $\tilde{x}$ to identify (m_1, σ^2) - which can also be pre-estimated separately as first-stage to reduce the number of parameters to be jointly estimated - as well as the mean and variance of the compensation and its covariance with the performance measure. Other moments to capture the shape of the compensation may include the skewness and kurtosis of the compensation, or the coefficients of a regression of compensation on the performance measure on a polynomial expansion.

5.3 Non-Parametric identification and Estimation with Exponential (CARA) Utilities

Margiotta and Miller (2000) and Gayle and Miller (2015) examine the identification of this model without restricting the form of the likelihood ratio as in (A2). They propose the following assumptions:

(A1') the agent has an exponential utility $u(w; a) = -e^{-r(w-ac)}$ with (in-the-utility) cost of effort $c > 0$ equal to c units of compensation;

discrete approximation is known to converge slowly to the true distribution.

(A2') a high enough performance measure reveals high effort, so that $f(x|0)/f(x|1)$ converges to zero as x becomes large;

(A3') the principal prefers to induce high effort,

$$\int f(x|1)(x - w(x))dx \geq \int f(x|0)(x - \underline{w})dx. \tag{5.10}$$

Assumption (A1') has an additional benefit. As shown in Fellingham, Newman and Suh (1985) and Margiotta and Miller (2000), any infinite-horizon contracting model in which the program is repeated and the principal and agent engage in a long-term contract, can be implemented as a sequence of single-period optimal contracts in which $w(x)$ represents the total change in agent wealth á la Core, Guay and Verrecchia (2003). Intuitively, the principal solves the agency problem in the current period by increasing agent wealth by $w(x)$. Then, because (i) implies that wealth scales all payoffs but does not affect the solution to the agency problem, the same program repeats in future periods.

Denoting $\mathcal{C} = e^{-rc}$ as the cost of effort as a proportional reduction in utility, one can rewrite (PC) as

$$\mathcal{C} \int f(x|1)e^{-r(w(x)-\underline{w})}dx = 1. \tag{PC'}$$

Further, equation (5.1) can be solved to express the unobserved p.d.f. $f(x|0)$ as a function of the exogenous and endogenous parameters of the model:

$$f(x|0) = f(x|1)\frac{r(\lambda + \mu)\mathcal{C}^{-1} - e^{rw(x)}}{\mu r}, \tag{5.11}$$

which, injecting in (IC'), implies that

$$\int f(x|1)e^{-rw(x)}dx = \frac{\mathcal{C}(\mu r e^{-r\underline{w}} + 1)}{r(\lambda + \mu)}. \tag{5.12}$$

The fact that a p.d.f. must integrate to one yields the following additional model restric-

tion:

$$\mu r = r(\lambda + \mu)\mathcal{C}^{-1} - \int f(x|1)e^{rw(x)}dx. \tag{D}$$

Using assumption (A2) and denoting $\overline{w}$ as the maximum wage (which can be estimated as a smooth fitted value of $w(.)$ for x large), one can evaluate (5.1) when x is large to obtain

$$r(\lambda\mathcal{C} + \mu) = e^{-r\overline{w}}. \tag{M}$$

Assumption (A3') is an inequality that, after substituting in $f(x|0)$ and given that $\int w(x)f(x|1)dx$ is increasing in risk-aversion r, set identifies that $r \leq \overline{r}$. Intuitively, the agent's risk-aversion cannot be so large that it would make inducing effort more costly than its potential effect on output. In summary, the model is identified up to a range of risk-aversion coefficients: the model has five unknown parameters $(\lambda, \mu, r, c, \underline{w})$ which can be set identified with four equations (PC'), (5.12), (D) and (M), as well as the inequality condition for (A3').

Gayle and Miller (2015) propose the following estimation procedure. Take risk-aversion as a known given and estimate all the other parameters of the model, as a function of r, from (PC'), (5.12), (D) and (M), using for example one of the approaches in Section 5.2. Then, evaluate the inequality (5.10) and drop the risk-aversion r if this inequality is violated. This procedure yields a set of acceptable risk-aversions and associated parameter values.

6 Structural Models of Disclosure Theory

6.1 Identification

Voluntary disclosure theory is a core pillar of accounting research. According to the theory, firms have verifiable information that, absent any friction, would be reported truth-

fully to investors: the unravelling principle, first used in Viscusi (1978), explains that investors would rationally price non-disclosing firms at the average information conditional on non-disclosure; in turn, implying that any non-disclosing firms above the average of all other non-disclosing firms would shift to disclose. In practice, such stark predictions are rarely observed, and disputes, in both the empirical and theoretical literature, continue about both the magnitude and the nature of disclosure frictions preventing this outcome.

In this section, we consider the empirical approach to estimating voluntary disclosure theory in Cheynel and Liu-Watts (2015). Their model shows that (i) frictions in the form of disclosure costs (Jovanovic 1982, Verrecchia 1983) can be identified and estimated without distributional knowledge of the distribution of manager's information, but (ii) frictions in the form of uncertain information endowment (Dye 1985, Jung and Kwon 1988) are only set identified.

While the model can be written with multiple periods, we develop here a single-period model. The firm privately receives an information $\tilde{s}$, where $\tilde{s} = \tilde{x}$ is the firm's expected cash flow with probability $q \in [0, 1]$ in which case we say that the firm is informed, or does not receive any information $\tilde{s} = "ND"$ with probability $1 - q$. Suppose that $\tilde{x}$ has a p.d.f. $f(.)$ and c.d.f. $F(.)$ with finite mean and support on $\mathbb{R}$. Then, the firm can make a disclosure $d(s) \in \{s, ND\}$, where $d(s) = s$ indicates that the verifiable information is reported truthfully, while $d(s) = ND$ indicates that the firm does not disclose information either because they were uninformed $s = ND$ or withhold $s = x$ strategically. The firm maximizes its price post disclosure decision but faces a cost $c(x) > 0$ when making a disclosure. The disclosure cost may be a function of x but, to reflect that higher x indicates better news from the perspective of the firm, assume that $x - c(x)$ is increasing (Bertomeu and Cianciaruso 2016).[17]

Denoting $P(d(s))$ as the market price given a disclosure $d(s)$, the firm discloses if

[17] Bertomeu and Cianciaruso (2016) show that this restriction is without loss of generality and, if it were violated, we could redefine another random variable such that $y - c(y)$ is increasing by reranking the states from the perspective of the decision-making firm.

$P(s) - c(s) \geq P(ND)$, i.e., its disclosure price is greater than its withholding price. To close the model, suppose that the firm is priced at its expected cash flow so that $P(x) = x$ and $P(ND) = \mathbb{E}(\tilde{x}|d(\tilde{s}) = ND)$ is priced at the expected information that would be withheld.

As is well-known in these models, the disclosure strategy reduces to a threshold τ such that firms disclose when their information $x \geq \tau$ is sufficiently favorable. A firm lying exactly at the threshold must be indifferent which (then) returns that

$$P(\tau) - c(\tau) = P(ND). \tag{6.1}$$

Two technical notes need to be emphasized at this point. First, the game may have multiple equilibria and, while it can be shown that equilibria with maximum price $P(ND)$ are Pareto dominant, no equilibrium selection is required for estimation purposes as long as firms with the same observables choose the same equilibrium. The equilibrium that is played can be generally identified from the sample and can be thought as an unknown parameter of the model if (6.1) does not have a unique solution. Second, the model restriction is a function of only $c(\tau)$, in the sense that any change in $c(x)$ that leaves $c(\tau)$ unchanged does not affect observed disclosure policies. So, hereafter, we shall be solely concerned in identifying $c \equiv c(\tau)$.

After noting that the withheld information is an expectation conditional on not receiving information or being informed and choosing to withhold:

$$P(ND) = \frac{qF(\tau)\mathbb{E}(\tilde{x}|\tilde{x} \leq \tau) + (1-q)\mathbb{E}(\tilde{x})}{qF(\tau) + (1-q)}. \tag{6.2}$$

Unfortunately, using (6.2) to identify $c(\tau)$ requires to know $P(ND)$ which, in (6.2), is a function of the unknown information endowment probability q. Hence, equation (6.2) adds one equation and one unknown and does not help recover the cost.

To address this, $P(ND)$ can be decomposed from the law of total probability as

$$\mathbb{E}(\tilde{x}) = q(1 - F(\tau))\mathbb{E}(\tilde{x}|d(\tilde{x}) = \tilde{x}) + ((1 - q) + qF(\tau))P(ND), \tag{6.3}$$

which, solving for $P(ND)$ and substituting in (6.1) yields the value of the cost function as in section 1.1,

$$c = \tau - \frac{\mathbb{E}(\tilde{x}) - q(1 - F(\tau))\mathbb{E}(\tilde{x}|d(\tilde{x}) \neq ND)}{(1 - q) + qF(\tau)}. \tag{6.4}$$

In equation (6.4), the cost is identified independently of knowledge of q as a function of the threshold τ, the probabilities of non-disclosure $(1 - q) + qF(\tau)$ and disclosure $q(1 - F(\tau))$, and the unconditional expectation $\mathbb{E}(\tilde{x})$. In the model, the threshold is the minimum disclosure while the probabilities of non-disclosure and disclosure can be recovered from the disclosure frequency. The last part requires the expected disclosure $\mathbb{E}(\tilde{x})$, which can be estimated either from data about (noisy) realizations of $\tilde{x}$ or, if an analyst consensus is available, by redefining the disclosure netting out the consensus so that $\mathbb{E}(\tilde{x}) = 0$ is defined in surprises. Vice-versa, it is readily seen that the cost is not identified if we do not have information to estimate $\mathbb{E}(\tilde{x})$. In what follows, we simplify notation using $\mathbb{E}(\tilde{x}) = 0$.

Although c can be point identified with this type of dataset, identification of uncertainty about information endowment is more problematic. To see why, note that we observe the frequency of non-disclosure as

$$f = (1 - q) + qF(\tau) \in (1 - q, 1), \tag{6.5}$$

but there are multiple combinations of the unobserved q and $F(\tau)$ consistent with f. Intuitively, observing a high probability of non-disclosure for given c can be caused by a high probability of not receiving information with, otherwise, a small probability of a

sufficiently bad type who choose not to disclose; or a low probability of not receiving information with a high probability of firms with x lying close to $\tau - c$.

Nevertheless, the probability of information endowment can be set identified. Noting first that $F(\tau) \in [0, 1]$, equation (6.5) implies the straightforward restriction that $q \geq 1 - f$ since, obviously, the probability of receiving information must be greater than the frequency of disclosure. Another bound can be derived by rewriting the indifference condition

$$\tau - c = \frac{qF(\tau)\mathbb{E}(\tilde{x}|\tilde{x} \leq \tau)}{f} \leq \frac{qF(\tau)\tau}{f} = \frac{(f - 1 + q)\tau}{f}, \tag{6.6}$$

where the last equality follows from $f = (1 - q) + qF(\tau)$. If $\tau < 0$,

$$q \leq \frac{\tau - cf}{\tau}. \tag{6.7}$$

To give more intuition, if the point-identified c is estimated to be negative, i.e., a benefit of disclosure, a point estimate $q = 1$ ruling out uncertainty about information endowment would imply full-disclosure, which is rejected by any sample with some withholding. Hence, the probability of being uninformed needed to rationalize the data must be bounded away from zero. As the frequency of non-disclosure f increases, the bound on the maximal probability that the manager is informed, increases as well.

6.2 Dynamic Disclosure Theory with Random Costs

The standard models of voluntary disclosure (Verrecchia 1983, Dye 1985) are static and therefore cannot fit dynamic aspects of disclosure behavior, such as the persistence of disclosure choices. Indeed, in practice, the more a firm has disclosed in the past, the more likely it is to continue disclosing in the future: firms form a reputation that endogenously bind them to continue disclosing (Graham, Harvey and Rajgopal 2005).

However, this explanation is unsatisfying if used as a complete theory of dynamic voluntary disclosure, because it neither explains what reputations are and why they can form. For example, if a firm faces a repetition of static disclosure problems, no reputation will form, implying that there must be economic primitives of the information process that are critical to reputations. Put differently, if, as is observed empirically, markets react more negatively when a firm ceases to disclose than absent prior disclosures, a complete explanation must explain what features of the information environment lead to this type of investor response. Structural models offer the opportunity to open the black box of reputations by estimating a dynamic theory of endogenous prices and voluntary disclosure. In other words, the structural model allows the researchers to interpret an empirical fact (persistence of disclosure) into a coherent theory of reputations.

A starting point to justify a structural model is to find empirical facts that call for a greater understanding of a theoretical mechanism. The left-hand panel of Figure 4 documents the probability of disclosure (i.e, releasing a management forecast in a given period) in prior and subsequent periods conditional on disclosure at date t. The right-hand side panel confirms that management forecasts appear more concentrated and tilted toward good news than actual earnings, thus suggesting that a model of strategic disclosure is appropriate.

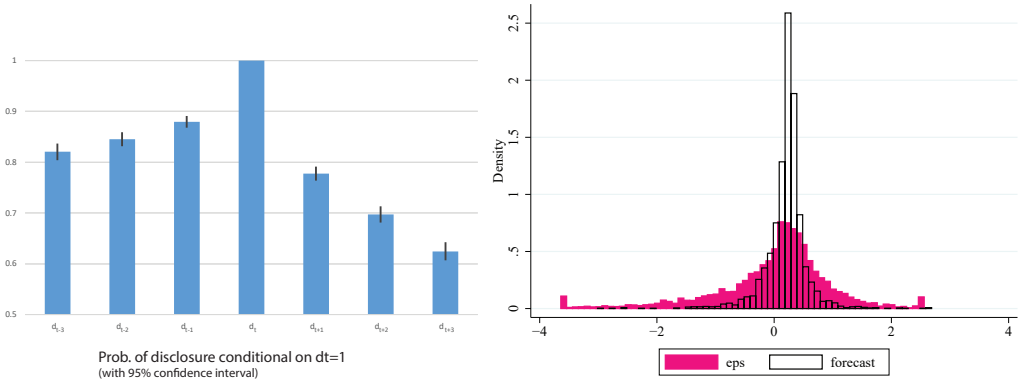


Figure 4:
Probability of Disclosure Conditional on Prior Disclosure

The left-hand panel reports disclosure frequencies conditional on $d_t = 1$, along with the 95% confidence interval. The right-hand panel reports the distribution of realized and forecasted EPS.

To make the voluntary disclosure model more realistic and amenable to empirical analysis, Bertomeu et al. (2020), hereafter BMM, consider a model where firms experience random disclosure costs, casting the analysis within a discrete choice setting (McFadden 1973, McFadden 1980) and modifying it to incorporate endogenous stock price incentives. In their model, the manager chooses disclosure $d \in \{0, 1\}$ to maximize a utility

$$u_t(d_t|x_t) = \alpha(d_t P_t(x_t) + (1 - d_t)P_t^{nd}) + d_t\beta + d_t\epsilon_t, \tag{6.8}$$

where β is a fixed disclosure benefit or cost depending on its signal and is a reduced-form for theories in which disclosures have real consequences.[18] The parameter α is the weight that a manager places on the firm's price; it can be influenced by factors such as executives' compensation package. The shock ϵ_t is a standard normal i.i.d. white noise, observed only by the manager which captures other time-varying factors affecting the disclosure decision. An expression for the prices $P_t(x_t)$ and P_t^{nd} is deferred until the solution of the disclosure problem is presented.

To introduce persistence in disclosure choices, BMM consider a setting where the manager's information endowment is persistent adapted from Einhorn and Ziv (2008). In each period, the manager's information state is represented by a serially correlated binary variable $\theta_t \in \{0, 1\}$ equal to one when the manager is informed and zero when the manager is uninformed. In formal terms, the manager's information endowment θ_t is a hidden Markov chain with a transition matrix

$$\Pi = \begin{pmatrix} 1 - \lambda_1 & \lambda_1 \\ \lambda_0 & 1 - \lambda_0 \end{pmatrix}, \tag{6.9}$$

[18] For example, disclosures may have implications on investment (Bertomeu and Cheynel 2015, Bertomeu, Cheynel and Cianciaruso 2021a), litigation (Caskey 2014, Marinovic and Varas 2016), endogenous certification (Marinovic and Sridhar 2015), cost of capital (Bertomeu, Beyer and Dye 2011, Cheynel 2013), as well as proprietary costs and product market coordination benefits (Cheynel and Ziv 2021, Bertomeu, Evans III, Feng and Tseng 2021e).

where $\lambda_0 \equiv \mathbb{P}(\theta_{t+1} = 0|\theta_t = 1) \in (0,1)$ denotes the probability of moving from the informed to the uninformed state, and $\lambda_1 \equiv \mathbb{P}(\theta_{t+1} = 1|\theta_t = 0) \in (0,1)$ denotes the probability of moving from the uninformed to the informed state. The information endowment is persistent when becoming uninformed is less likely than remaining uninformed, or $\lambda_0 < 1 - \lambda_1$. As in Dye (1985), when informed, the manager chooses whether to disclose, choosing $d_t \in \{0,1\}$ to maximize (6.8) where $P_t(x_t) = \mathbb{E}(x_t|d_t = 1, x_t)$ is the price conditional on disclosure and $P_t^{nd} = \mathbb{E}_t(x_t|d_t = 0)$ is the non-disclosure price conditional on all prior public information.

The distribution of forecast and earnings surprises is specified as

$$\begin{pmatrix} x_t \\ e_t \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_e^2 \end{bmatrix} \right), \tag{6.10}$$

where $cov(x_t, e_t) = cov(x_t, x_t + v_t) = Var(x_t) = \sigma_x^2$ captures the amount of private information that the manager may know in advance.

Estimating the model requires to compute the expected price consequence of a non-disclosure. Unlike in a standard disclosure model, investors do not observe ϵ_t and therefore must price the firm in response to a random threshold. BMM show that the non-disclosure price P_t^{nd} must be the unique solution to the fixed point $\Gamma(P_t^{nd}) = P_t^{nd}$ with

$$\Gamma(y) = \frac{1 - p_t}{p_t} \int x\Phi(-\alpha x - \beta)\frac{1}{\sigma_x}\phi(\frac{x+y}{\sigma_x})dx, \tag{6.11}$$

where $\Phi(.)$ and $\phi(\cdot)$ are the c.d.f. and p.d.f. of the standard normal distribution respectively, which yields the non-disclosure price as a function of current investor beliefs $p_t = 1 - \mathbb{E}_t(\theta_t)$ that the firm is uninformed.

From a computational standpoint, one can solve for the non-disclosure price using (6.11) for any p_t; of course, p_t is not directly observable, but, as will be shown later on, can be derived recursively from a sequence of prior disclosures and earnings using Bayes

rule. This also creates a computational issue: given that p_t varies, one would have to solve the fixed point problem separately for each observation and would impractically require solving n fixed point for a sample of n observations. Fortunately, this problem can be addressed numerically by solving the fixed point on a grid of possible values of p_t in $[0, 1]$ and interpolating all remaining values. BMM show that a 15-point spline interpolation of the fixed point problem is nearly indistinguishable from the pointwise solution of the fixed point.

BMM implement an MLE approach by computing the likelihood at a disclosure or non-disclosure at each period. The likelihood $L_t^{d_t}(d_t x_t, e_t)$ of an observation in period t is a function of two possible outcomes:

(i) Conditional on a disclosure, the distribution of the forecast is $x_t|e_t \sim N(\frac{\sigma_x^2}{\sigma_e^2}e_t, \sigma_x^2(1-\frac{\sigma_x^2}{\sigma_e^2}))$, which yields a likelihood

$$L_t^1(x_t, e_t) = \frac{(1-p_t)\Pr(\epsilon_t \geq \alpha(P_t^{nd} - x_t) - \beta|e_t, x_t)}{\sigma_x\sqrt{1-\frac{\sigma_x^2}{\sigma_e^2}}}\phi(\frac{x_t - \frac{\sigma_x^2}{\sigma_e^2}e_t}{\sigma_x\sqrt{1-\frac{\sigma_x^2}{\sigma_e^2}}})\frac{1}{\sigma_e}\phi(\frac{e_t}{\sigma_e}), \tag{6.12}$$

and the probability of being informed is updated to $p_{t+1} = \Pr(\theta_{t+1} = 0|\theta_t = 1) = \lambda_0$;

(ii) Conditional on non-disclosure,

$$L_t^0(0, e_t) = \underbrace{(p_t + (1-p_t)\Pr(\alpha x_t + \epsilon_t < \alpha P_t^{nd} - \beta|e_t))}_{\equiv \nu_t}\frac{1}{\sigma_e}\phi(\frac{e_t}{\sigma_e}). \tag{6.13}$$

Then, investors update the probability that the manager will be informed in the next period using Bayes rule:

$$p_{t+1} = \frac{p_t}{\nu_t}\lambda_1 + (1 - \frac{p_t}{\nu_t})(1 - \lambda_0). \tag{6.14}$$

We can now write the (log) likelihood of a time-series of disclosures

$$\mathcal{L}((d_t, d_t x_t, e_t)_{t=1}^{T}) = \sum_{t=1}^{T} \ln L_t^{d_t}(d_t x_t, e_t), \tag{6.15}$$

where p_t is updated recursively using equation (6.14).

The last step of the analysis prior to conducting the estimation is to explain what information in the sample is captured by the estimation methods to identify the parameters of interest. For simplicity, we focus here only on price motives and plot the density of the disclosures (x) in the model for various values of α in Figure 5. At one extreme, when $\alpha = 0$, the probability of disclosure is $(1-p_t)\Phi(\beta)$; the manager ignores the effect of disclosure on price; and, conditional on being informed, disclosure is driven purely by the preference shocks ϵ_t; hence the distribution is symmetric around zero. As α increases, the model predicts selection over favorable disclosures in the form of a truncated normal distribution coexisting with disclosures driven by non-price incentives, $\beta + \epsilon_t$. Thus, skewness in the distribution of disclosures serves as the identifying variation for recovering α. Because MLE uses information from the shape of the distribution of disclosures, the method is well-suited to capture this identifying feature.

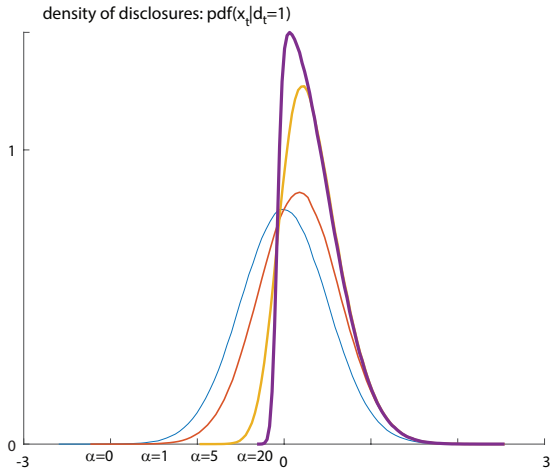


Figure 5:
Likelihood of Disclosure across Parameter Regions

6.3 Dynamic Disclosure Theory with Forward-Looking Preferences

The previous model assumes that the manager maximizes current stock prices in a myopic manner, as if disclosure decisions did not have implications for future prices. However, in the presence of persistent (information endowment) shocks, a disclosure today has an effect on the manager's future disclosure behavior, and thus on the future non-disclosure prices. This reputation concern affects the manager's current disclosure choice.

To capture the dynamic nature of disclosure choices, Bertomeu, Marinovic, Terry and Varas (2022b), hereafter BMTV, estimate a model in which a manager sells shares over time. Following Benmelech, Kandel and Veronesi (2010) and Beyer and Dye (2012), managers maximize the discounted value of a firm's stock price with expected utility in period t

$$\mathcal{U}_t = \mathbb{E}_t \left(\sum_{k=t}^{\infty} \beta^{k-t} P_k \right). \tag{6.16}$$

Here, $\beta \in (0, 1)$ is a subjective discount factor interpretable as the rate at which managers sell shares or are exposed to future share prices via compensation vesting schedules (Edmans, Fang and Lewellen 2017, Marinovic and Varas 2019), P_k is the firm's market price, and $\mathbb{E}_t(.)$ is the expectation at the beginning of period t. Apart from the addition of forward-looking concerns, the model in 6.2 is simplified so that there is no fixed or random cost of disclosure.

While Bertomeu et al. (2020) net out the analyst consensus to express forecasts and realized earnings in pre-announcement surprises, the process for earnings, forecasts and earnings is made more explicit in BMTV so that it can be jointly estimated. At the start of the period, the market observes a consensus analyst forecast $\hat{c}_t$ about end-of-period earnings e_t. Then, the manager may receive additional information in the form of a private signal $\hat{s}_t$, a signal observed if the indicator variable $\theta_t = 1$. Conversely, if $\theta_t = 0$, the manager does not receive additional material information. The market does not know the

realization of θ_t but forms expectations about managers' information given by a probability $p_t = 1 - \mathbb{E}_t(\theta_t)$. If information is received, the manager may decide to voluntarily disclose $\hat{s}_t$, i.e., issue a forecast $d_t = \hat{s}_t$ about end-of-period earnings e_t. By convention, $d_t = ND$ indicates that no forecast is made. Then, the price P_t forms, reflecting the value of future earnings discounted at an objective mrate of return r and conditional on all public information $\mathcal{H}_{t-1}$ up to the end of period $t-1$ as well as any new information $\{\hat{c}_t, d_t\}$,

$$P_t = \mathbb{E}\left(\sum_{k=t}^{\infty} \frac{e_k}{(1+r)^{k-t}} | \mathcal{H}_{t-1}, \hat{c}_t, d_t\right). \tag{6.17}$$

At the end of the period, the firm releases its earnings e_t. These earnings are publicly observed and the public information set is updated to $\mathcal{H}_t = \{\mathcal{H}_{t-1}, \hat{c}_t, d_t, e_t\}$.

The process of earnings, the consensus analyst signal, and the manager's signal jointly satisfy (i) $e_t = \rho e_{t-1} + u_t$, (ii) $\hat{c}_t = e_t + v_t$ and (iii) $\hat{s}_t = e_t + w_t$, where $\varepsilon_t = (u_t, v_t, w_t)'$ is an iid normal vector with variance-covariance matrix $diag(\tau_u, \tau_v, \tau_w)^{-1}$. Earnings e_t follow an AR(1) process, and information observed by investors and managers $\hat{c}_t$ and $\hat{s}_t$ are noisy orthogonal signals. The model's predictions are invariant to means, so without loss of generality one can normalize all processes and signals to mean zero.

While one can state the model in terms of a vector of signals $(\hat{c}_t, \hat{s}_t)$, empirically, analysts and managers seen to communicate in terms of their expectations about future earnings. Hence, it is useful to normalize the variables to posterior expectations: $c_t = \mathbb{E}(e_t | \hat{c}_t, \mathcal{H}_{t-1})$ for the analyst consensus forecast and $s_t = \mathbb{E}(e_t | \hat{s}_t, \hat{c}_t, \mathcal{H}_{t-1})$ for the manager's forecast. When restating the model in terms of posterior expectations, the joint stochastic process of earnings and expectations becomes

$$\begin{pmatrix} e_t \\ c_t \\ s_t \end{pmatrix} = \rho \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} e_{t-1} + \begin{pmatrix} 1 \\ \frac{\tau_v}{\tau_u+\tau_v} \\ \frac{\tau_v+\tau_w}{\tau_u+\tau_v+\tau_w} \end{pmatrix} u_t + \begin{pmatrix} 0 \\ \frac{\tau_v}{\tau_u+\tau_v} \\ \frac{\tau_v}{\tau_u+\tau_v+\tau_w} \end{pmatrix} v_t + \begin{pmatrix} 0 \\ 0 \\ \frac{\tau_w}{\tau_u+\tau_v+\tau_w} \end{pmatrix} w_t. \tag{6.18}$$

We write next the pricing function conditional on disclosure and non-disclosure. As a first step, let us consider the pricing function given by a threshold disclosure strategy $s \geq \tau(p, s)$. Letting $P^D(z, s)$ and $P^{ND}(p, z)$ be the market prices conditional upon disclosure and non-disclosure, respectively,

$$P^D(z, s) = \frac{1+r}{1+r-\rho}\mathbb{E}(e|z, s), \tag{6.19}$$

$$P^{ND}(p, z) = \frac{1+r}{1+r-\rho}\frac{p\mathbb{E}(e|z) + (1-p)\mathbb{E}(e\mathbf{1}_{s\leq\tau(p,s)}|z)}{p + (1-p)\mathbb{E}(\mathbf{1}_{s\leq\tau(p,s)}|z)}, \tag{6.20}$$

where r is the discount rate so that the ratio $\frac{1+r}{1+r-\rho}$ represents the valuation multiplier on current expected earnings given persistence ρ. In (6.20), $P^{ND}(p, z)$ is a weighted average between the payoff if the manager is uninformed $\mathbb{E}(e|z)$ and if the manager is informed but strategically withholding $\mathbb{E}(e\mathbf{1}_{D^c(p,z)}|z)$.

The updated probability p' that the manager will be uninformed in the next period is then obtained

$$p' = \varphi(p, z, e) \equiv \frac{p(1-\lambda_1) + (1-p)\lambda_0\mathbb{E}(\mathbf{1}_{s\leq\tau(p,s)}|e)}{p + (1-p)\mathbb{E}(\mathbf{1}_{s\leq\tau(p,s)}|e)}. \tag{6.21}$$

What remains to be determined is the manager's strategy $\tau(p, s)$. To obtain intuition for this problem, consider the simpler problem in which the manager is myopic and maximizes current price ($\beta \to 0$). Then, the disclosure threshold is characterized by the indifference condition $P^D(z, \tau(p, s)) = P^{ND}(p, z)$ and coincides with the static solution in Jung and Kwon (1988). When the manager cares about future prices, an additional effect occurs because, by disclosing, the informed manager loses their information advantage and makes investors more skeptical in future periods. Thus, BMTV show that the disclosure threshold increases when the manager cares more about future prices.

The formal characterization of the threshold requires to write the manager's value function from current and future prices. We can define the manager's optimization prob-

lem in terms of four state-contingent value functions:

$$V_1(p,z,s) = \max_{d\in\{0,1\}} P^{j(d)}(s,z) + \beta\mathbb{E}\Big[(1-\lambda_0)V_1(\lambda_0,z',s') + \lambda_0 V_0(\lambda_0,z')\Big|z,s\Big] \quad (6.22)$$

$$V_0(p,z) = P^{ND}(p,z) + \beta\mathbb{E}\Big[\lambda_1 V_1(p',z',s') + (1-\lambda_1)V_0(p',z')\Big|z\Big], \quad (6.23)$$

where primes denote next-period values, $j(d) = D$ if $d = 1$ and D otherwise, $V_1(p,z,s)$ is the informed manager's value function prior to making a disclosure choice, and $V_0(p,z)$ is the value function of an uninformed manager.

Swapping current prices in the myopic model with value functions conditional on each disclosure choice, the disclosure threshold is set at the indifference condition that equates the payoff from disclosure versus non-disclosure (6.22) :[19]

$$V_1^D(p,z,\tau(p,z)) = V_1^{ND}(p,z,\tau(p,z)). \quad (6.24)$$

The computational strategy to numerically solve this model will serve to illustrate the use of policy function iteration, which offers a usually more efficient alternative to value function iteration. The steps are as follows:

1. Discretize the state space.

2. On the s-th iteration of the solution algorithm, guess a disclosure policy $d^{(s)}(p,e,c,s)$.

   (a) Assume that market beliefs and manager actions are governed by $d^{(s)}$, and iterate forward on the system of Bellman equations above until the implied $V_1^{(s)}$, $V_0^{(s)}$ converge to some tolerance.

   (b) Compute the stationary distribution $\mu_1^{(s)}(p,e_{-1},c,s)$ and $\mu_0^{(s)}(p,e_{-1},c)$ of the model given $d^{(s)}$, as well as the exogenous distributions in the model. This

[19] A natural question is whether a disclosure threshold exists in this type of problem, so the approach used in BMTV is more general and allows for disclosure strategies that need not follow a threshold equilibrium. Grubb (2011) shows that if the manager is perfectly informed about future earnings, this model has indeed no equilibrium; empirically, BMTV, however, show that the estimates are consistent with a threshold equilibrium because forecasts include a significant amount of noise on future earnings.

involves repeatedly pushing forward weight on a histogram given the policies and exogenous transitions until the distributions stabilize to within some tolerance.

(c) Compute a *new* policy $d^{(s+1)}(p, e_{-1}, c, s)$, simply given by the maximization in (6.22) using $V_1^{(s)}$ and $V_0^{(s)}$ in lieu of the true V_1 and V_0.

(d) Then, compute an error measure given by the mean absolute difference between $d^{(s+1)}$ and $d^{(s)}$, weighted by the ergodic distributions $\mu_1^{(s)}$ and $\mu_0^{(s)}$. This error is exactly equal to the probability of disclosure policy deviation given assumed market beliefs. When this error is sufficiently small, we have computed an equilibrium.

3. Once the model is solved, one can simulate and compute moments as desired for input into the structural estimation routine.

The discretization of the exogenous processes for earnings, consensus forecasts, and manager signals uses the Tauchen (1986) method. The remaining numerical approach to the discrete-state dynamic programming problem follows the methods outlined in Judd (1998b). The two parameters relating to time preferences β and discounting r, are not well identified by a theory of dynamic disclosure and instead recovered from external information sources, e.g., equity pay duration in Gopalan, Milbourn, Song and Thakor (2014), and the discount rate used in prior literature.

The remaining parameters are estimated with two-step simulated method of moments (SMM) following Bazdresch et al. (2018). In step 1, the exogenous processes relating to earnings and analyst consensus forecasts are directly estimated from their time-series properties. In step 2, the remaining parameters governing the evolution and precision of managers' information are obtained by matching moments in the sample with simulated data from the model. The data set has about 5,000 firms with an average of five fiscal years for each: matching the time structure of this panel, the simulation includes 20,000

firms for five years each in our model after discarding an initial burn-in period for each firm.[20]

The unconditional probability of disclosure, as well as the persistence of disclosure over one, two, and three periods, help identify the information switching probabilities λ_0 and λ_1. Because the persistence and precision of information, linked to each of the parameters λ_0, λ_1, and σ_w, determine the extent to which market inference reacts to disclosure, and because this market reaction determines the endogenous selection into disclosure by managers, another natural moment to target is the difference between the average level of earnings when disclosing versus unconditionally. In summary, the moments help identify the persistence of the information endowment, and the quality of the information received by the manager.

6.4 Unverifiable Disclosure

Reported earnings are often manipulated by managers, both within GAAP and in reports supplementary to GAAP (McClure and Zakolyukina 2022). But the propensity of firms to manipulate reports may be unknown to investors. Investors hold beliefs about the firm's propensity to report truthfully, i.e, the firm's credibility, and update those beliefs over time based on the firm's reporting behavior.

Bertomeu and Marinovic (2016) assume that a manager may be able to misreport the earnings number, possibly because of failures in internal controls. Specifically, with probability γ, managers truthfully report the firm's true earnings $x_t \sim N(0, \sigma_x^2)$ and, with complementary probability, they lie to maximize the firm's current stock price P_t. As this behavior repeats over time, the manager loses credibility, and the market becomes less responsive to the manager's report. Marinovic (2013) and Liang, Marinovic and

[20] Our estimation requires attention to this sort of detail. To account for permanent firm heterogeneity unrelated to earnings innovations, we target the autocorrelation and standard deviation of earnings after removing firm fixed effects. So we must match the empirical implications of removing fixed effects in short samples within our simulated data to ensure appropriate inference about the earnings process.

Varas (2018) propose empirical implementations of this learning process, to estimate the distribution of truthfulness from firms' reporting choices. A simplified version of their model is developed below.

The firm's true value is a linear function of cumulative earnings X_t. The manager observes x_t and makes a cumulative report R_t to maximize current price - for example, we may interpret x_t as earnings and R_t as the book value of equity. Marinovic (2013) shows that there exists a constant K_t such that the price is given by

$$P_t = \beta \min(R_t, K_t). \tag{6.25}$$

Intuitively, the market views reports below R_t as credible but fully discounts any report above K_t. Intuition for a flattened pricing function can be obtained by contradiction. If the pricing function were strictly increasing (or, more generally, if it attained a maximum at one value), all untruthful managers would choose the highest price report. But then, such report would be fully-revealing and lack any credibility. To avoid this, the equilibrium must be such that untruthful managers use a signal jamming strategy by reporting on the interval randomly $[K_t, \infty)$, which in turn requires them to be indifferent and achieve a flat price.[21]

Since investors use Bayes rule to price reports above K_t, it must be that K_t is equal to the conditional expectation of drawing a truthful manager with true cumulative earnings above K_t or an untruthful manager whose cumulative earnings are just the unconditional mean of zero:

$$K_t = \frac{\gamma_{t-1}(1 - F_t(K_t))\mathbb{E}(X_t | X_t \geq K_t)}{\gamma_{t-1}(1 - F_t(K_t)) + 1 - \gamma_{t-1}}, \tag{6.26}$$

where γ_{t-1} is the probability the manager is truthful conditional on a past history of disclosures and F_t is the c.d.f. of cumulative true earnings $X_t = \sum_{k=0}^{T} x_k$. It can be shown

[21] As an alternative to mixed strategies, Bertomeu and Marinovic (2016) show that an equivalent price schedule can be implemented under pure strategies if truthful firms may under-report to credibly convey their information.

that this equation has a unique solution.

Having solved for K_t, the same logic as (6.26) can be applied pointwise to any report $R_t \geq K_t$ to derive the randomization strategy of the manager, since the conditional expectation of the true cash flow must equal K_t. Denote $\sigma_t(R)$ as the p.d.f. of the manager's reporting strategy,

$$K_t = \frac{\gamma_{t-1} F_t'(R_t) R_t}{\gamma_{t-1} F_t'(R_t) + (1-\gamma_{t-1})\sigma_t(R_t)}. \tag{6.27}$$

The above equation can be solved for $\sigma_t(R_t)$ to recover the untruthful's manager reporting strategy: for any $R_t \geq K_t$,

$$\frac{\sigma_t(R_t)}{F_t'(R_t)} = \frac{\gamma_{t-1}}{1-\gamma_{t-1}} \frac{R_t - K_t}{K_t}. \tag{6.28}$$

Note that, as is intuitive, the likelihood ratio of untruthful to truthful managers decreases when observing more favorable reports and converges to zero when attaining the lower bound K_t. As in disclosure models, one can then recursively update the assessed probability of being untruthful in each period:

$$\gamma_{t+1} = \frac{\gamma_t F_t'(R_t)}{\gamma_t F_t'(R_t) + (1-\gamma_t)\sigma_t(R_t)}. \tag{6.29}$$

This model can be estimated via MLE, by matching, at a firm level, the theoretical distribution of reported earnings to its empirical analogue. Below is a sketch of the algorithm one can use to build the likelihood function. For a given set of parameters (σ_x, γ), one can compute K_1. A report R_1 has thus a likelihood $\gamma F_1'(R_1) + (1-\gamma)\sigma_1(R_1)\mathbf{1}_{R_1 > K_1}$. Based on the realized report R_1, one can update γ to γ_1 using Bayes' rule in (6.29) and compute K_2 and $\sigma_2(.)$. A report R_2 has thus likelihood $\gamma_1 F_2'(R_2) + (1-\gamma_1)\sigma_2(R_2)\mathbf{1}_{R_2 > K_2}$. The process repeats iteratively for all periods $t = 1, 2, .., , T$ in a career.

7 Structural Models of Earnings Management

7.1 Static Price Incentives

In traditional earnings management theory, the manager trades off the cost of increasing an earnings report against the potential increase in price. Bertomeu, Cheynel, Li and Liang (2021b), hereafter BCLL, develop a closed-form estimation procedure to estimate an implied dollar-value estimate of earnings management. The manager of the firm privately observes a fundamental signal about the true earnings of the firm x, which is drawn from a distribution with full support on $\mathbb{R}$, p.d.f. $f(.)$, and c.d.f. $F(.)$.

Let $R(x) \in \mathbb{R}$ denote the manager's reporting strategy, and $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ denote a mapping that associates a price $\gamma(r)$ to each report r. For any *conjectured* increasing reporting strategy $\overline{R}(x)$ and observed report r, investors respond to earnings with

$$\gamma(r) = \mathbb{E}(\alpha(\tilde{x})|\overline{R}(\tilde{x}) = r),$$

where $\alpha(.)$ is a continuous function representing the mapping between unmanaged earnings and the firm value.

The manager maximizes the market reaction net of costs:

$$R(x) \in argmax_r \quad \overline{\gamma}(r) - \frac{1}{\theta}\psi(r - x),$$

where $\psi(.)$ is a twice-differentiable convex function with $\psi(0) = \psi'(0) = 0$ and $1/\theta$ is the earnings management cost.[22]

A fully-separating signalling equilibrium is defined as an increasing reporting function $R(x)$ and market reaction $\gamma(x)$ such that:

[22]This model has the single-crossing property of standard Spence signalling, which selects the full-separating equilibrium when it exists under traditional refinements (Cho and Kreps 1987). Further, pooling equilibria, if selected, predict holes in the support of the distribution of reports (Guttman, Kadan and Kandel 2006).

(i) investors price the firm according to correct equilibrium conjectures, that is, equation (7.1) is satisfied at $\overline{R}(x) = R(x)$, whenever possible;

(ii) managers optimally select their reporting strategy, that is, equation (7.1) is satisfied for any x at $\overline{\gamma}(x) = \gamma(x)$.

In an equilibrium, the first-order condition on the manager's problem in (7.1) is

$$\gamma'(R(x)) = \frac{1}{\theta}\psi'(R(x) - x). \tag{7.1}$$

Equation (7.1) states that the marginal benefit of earnings management $\gamma'(.)$, on the left-hand side, must equal the cost $\frac{1}{\theta}\psi'(R(x)-x)$, on the right-hand side. This expression is intuitive: given any earnings realization x, the equilibrium bias $R(x)-x$ will be higher if the market response $\gamma(.)$ is more sensitive to reported earnings.

For instance, if we set a quadratic cost of earnings management $\psi(x) = x^2$, the bias for a given earnings report

$$R(x) - x = \frac{\theta}{2}\gamma'(R(x)) \tag{7.2}$$

is proportional to the slope of the relation between earnings and price. If $\gamma(.)$ is linear and the support of x is unbounded from below, then the bias becomes solely a function of $1/\theta$ and is constant in reported earnings (Dye 1988, Stein 1989, Einhorn and Ziv 2012).

Equation (7.1) identifies earnings management up to a proportionality factor θ. To identify the levels of earnings management, some a priori knowledge of the shape of unmanaged earnings comes into play. If we make a guess about θ, we can reconstruct from Equation (7.1) a predicted distribution of reported earnings. A good choice of θ should then recover a distribution of reported earnings in the model that is as close as possible to the distribution of reported earnings observed empirically.

Formally, let the distribution of reported earnings $r = R(x)$ have p.d.f. $g(r)$. The

density of a random variable (r) that is a function of another (x) can be expressed as

$$g(r) \quad = \quad \frac{1}{|R'(R^{-1}(r))|} f(R^{-1}(r)), \tag{7.3}$$

where, in this type of model, a separating equilibrium implies that the reporting strategy is increasing and differentiable so that we can rewrite $|R'(R^{-1}(r))| = R'(R^{-1}(r))$.

Applying the implicit function theorem on (7.1) to obtain $R'(.)$,

$$R'(x) = -\frac{\psi''(R(x) - x)}{\theta\gamma''(R(x)) - \psi''(R(x) - x)}. \tag{7.4}$$

Substituting (7.3) into (7.4), the model-implied likelihood of the accounting reports is

$$g(r) = \frac{\theta^{-1}\psi''(r - R^{-1}(r)) - \gamma''(r)}{\theta^{-1}\psi''(r - R^{-1}(r))} f(R^{-1}(r)). \tag{7.5}$$

Equation (7.5) expresses the partial likelihood of the model in closed-form and suggests a natural estimation procedure by maximizing the log-likelihood $\frac{1}{n}\sum_i \ln g(r_i)$ for any sample $(r_i)_{i=1}^n$. However, this log-likelihood is not observable so several steps are required to obtain a feasible estimate of $\hat{g}(.)$. Bertomeu et al. (2021b) assume that earnings management costs are quadratic with, for any bias b, a cost $\psi(b) = b^2$ and the distribution of unmanaged earnings is normally distributed with mean m_x and standard deviation σ_x, with a p.d.f. function $\phi(x)$.

The mapping $\gamma(.)$ between earnings and prices and the bias $b(r) = r - R^{-1}(r)$ in (7.4) are not directly observable. However, $\gamma(.)$ can be estimated from any consistent non-parametric fitting procedure of observed prices on reported earnings, to obtain $\hat{\gamma}(.)$ and numerically differentiating the estimated function twice yields $\hat{\gamma}'(.)$ and $\hat{\gamma}''(.)$ in (7.5).

With these assumptions, the reporting bias is given from (7.2) as

$$b(r) = (\psi')^{-1}(\theta\gamma'(r)) = \frac{1}{2}\theta\gamma'(r). \tag{7.6}$$

Plugging in $b(r)$ in (7.5), jointly with the normality of x and the quadratic earnings management cost, implies an estimated likelihood

$$\hat{g}(r;\theta,m_x,\sigma_x) \equiv (1-\hat{\gamma}''(r)\frac{\theta}{2})\frac{1}{\sigma_x}\phi(\frac{1}{\sigma_x}(r-\frac{\theta}{2}\hat{\gamma}'(r)-m_x)), \tag{7.7}$$

where $\phi(.)$ is the p.d.f. of the standard normal distribution.

BCLL use a smoothing spline to fit stock price responses on earnings and recover $\hat{\gamma}$ and its derivatives $\hat{\gamma}'$. Alternatively, one can also use a polynomial function to estimate the price responses on earnings.[23] Given the estimates of $\hat{\gamma}$ and $\hat{\gamma}'$, the model simplifies to the following estimated likelihood function for a given sample of earnings $(r_i)_{i=1}^n$,

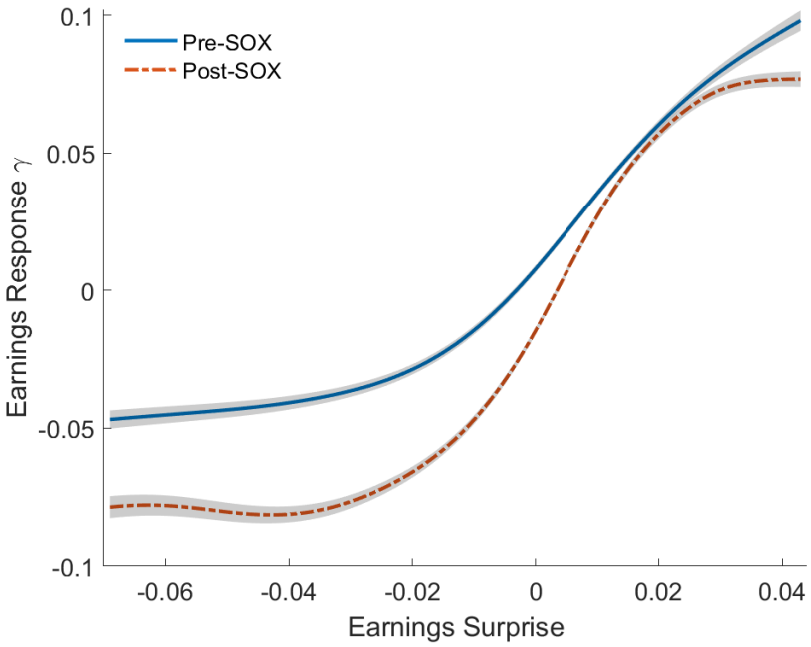
$$Lik(\theta,m_x,\sigma_x) \equiv \prod_{i=1}^{n}\frac{1}{\sigma_x}(1-\hat{\gamma}''(r_i)\frac{\theta}{2})\phi(\frac{1}{\sigma_x}(r_i-\frac{\theta}{2}\hat{\gamma}'(r_i)-m_x)),$$

where $\phi(.)$ is the p.d.f. of the standard normal. The parameters are the earnings management cost θ, the mean of unmanaged earnings m_x and the standard deviation of the unmanaged earnings σ_x.

Using earnings surprises as a proxy for reported earnings and cumulative abnormal stock return as proxy for price response, an estimate the market response function γ in plotted in Figure 6.

[23] For example, Bird, Karolyi and Ruchti (2019) define stock returns as a k-order polynomial function of earnings surprise and an additional j-order polynomial function of earnings surprise if the manager meets or beats analysts' forecast.

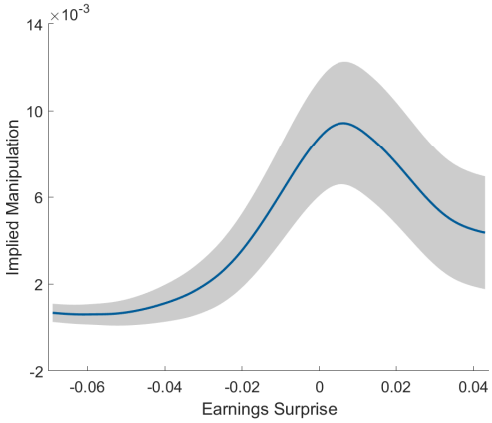
Figure 6: Earnings Response Functions



Estimation results in BCLL show that earnings management (the difference between reported earnings and unmanaged earnings) is on average 0.6% of the equity, which is modest in general. The cost of earnings management has increased by about 25% percent after SOX, implying that the post-regulatory environment made it more difficult to manage earnings. However, the average earnings management only slightly decreases, because the benefit from the capital market increased in the post period. With the estimates, we can also recover the implied reporting bias by: $\hat{R}(x) - x = \frac{\theta}{2}\hat{\gamma}'(\hat{R}(x))$:

Figure 7: Implied Earnings Management

Panel A: Pre-SOX Period



Panel B: Post-SOX Period

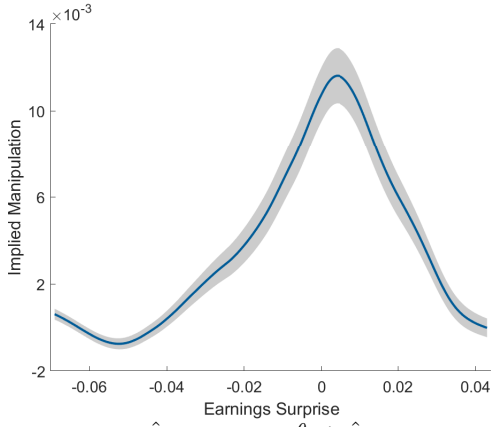


Figure 4 from BCLL. Implied earnings management is obtained by $\hat{R}(x) - x = \frac{\theta}{2}\hat{\gamma}'(\hat{R}(x))$, where θ is estimated using maximum likelihood estimation. Panel A presents the implied earnings management in the pre period (1990-2001) and Panel B presents the implied earnings management in the post period (2003 - 2014). The shaded areas are confidence intervals for one standard deviation. Both Earnings surprise and implied earnings management are scaled by the beginning-of-the-year book value of equity.

While in BCLL, investors can extract the bias from reported earnings, this approach can be generalized to models as in Fischer and Verrecchia (2000) in which the additional noise contaminates investors' inference process. Bertomeu, Li, Cheynel and Liang (2022a) expand on this model assuming that the

$$R(x, y) \in argmax_r \quad y\overline{\gamma}(r) - \frac{1}{\theta}\psi(r - x), \tag{7.8}$$

where $\tilde{y}$ represents a random incentive privately observed by the manager and drawn from a normal distribution $N(m_y, \sigma_y^2)$, implying a revised bias

$$b(r) = (\psi')^{-1}(\theta\gamma'(r)) = \frac{1}{2}y\theta\gamma'(r). \tag{7.9}$$

Unfortunately, this bias is no useable to compute the likelihood since it depends on y that is unobservable to the econometrician. To estimate the model, Bertomeu et al. (2022a)

rely on additional information from observed misstatements (v_i, z_i), where $v_i \in \{0, 1\}$ is a binary variable indicating the occurrence of a misstatement and $v_i z_i \in \mathbb{R}$ is a detected misstatement. The probability of detecting the misstatement is assumed to be a function $\zeta(b(r_i))$.

Under these assumptions, for a given set of parameters, one can simulate a distribution of biases conditional on a report r_i from (7.9), which implies a likelihood of detection by taking the expectation $\mathbb{E}(\zeta(b(r_i)))$ in $\tilde{z}$ and an implied distribution of misstatements. Parameter estimates can then be obtained by minimizing the difference between the joint distribution of detected misstatements and earnings in the data, and its model analogue.

7.2 Dynamic Price Incentives

Building on Holmström (1999), Beyer et al. (2019) (hereafter BGM) estimate a model of misreporting model where, as in Dye and Sridhar (2008), GAAP earnings introduce random noise. Consider below a simplified version of BGM. The firm's unknown fundamental is θ_t and evolves as an AR(1):

$$\theta_t = \rho\theta_{t-1} + \varepsilon_t, \tag{7.10}$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ is white noise. The manager privately observes true earnings $e_t = \theta_t$, equal to the fundamental. However, the market only observes the manager's report r_t, which may be manipulated. The manager's payoff in period t is given by

$$p_t - \frac{c}{2}(r_t - e_t - \zeta_t)^2, \tag{7.11}$$

where $p_t = \beta E(\theta_t | r_1, r_2, ..., r_t) \equiv \beta E_t(\theta_t)$ and $\zeta_t \sim N(0, \sigma_\zeta^2)$ is another white noise term capturing random differences between the report fully compliant with GAAP $e_t + \zeta_t$ and which would imply a cost of zero, and the true economic fundamental e_t. Conjecturing a

linear reporting strategy, by Bayes' rule, the price must evolve as

$$p_t = p_{t-1} + \beta\gamma_t(r_t - \mathbb{E}_{t-1}(r_t)), \tag{7.12}$$

where γ_t is an endogenous coefficient which may depend on a history of past reports. Maximizing the manager's objective in equation (7.11) leads to the equilibrium strategy

$$r_t = \frac{\beta\gamma_t}{c} + e_t + \zeta_t. \tag{7.13}$$

Since the price response γ_t is a time-varying but deterministic variable, the market observes a noisy signal of the fundamentals. We focus on the steady state of this model where, after enough periods, the new uncertainty due to the innovation in fundamentals ε_t is exactly offset by learning from the report.[24] Formally, denoting σ_θ^2 as the steady-state posterior variance of θ_t, the price response coefficient is constant and given by

$$\gamma_t = \frac{Cov_{t-1}(r_t, \theta_t)}{Var_{t-1}(r_r)} = \frac{\rho^2\sigma_\theta^2 + \sigma_\varepsilon^2}{\rho^2\sigma_\theta^2 + \sigma_\varepsilon + \sigma_\zeta^2}. \tag{7.14}$$

To characterize the steady-state posterior variance of θ_t implied by (7.14), one may note that forming an expectation for θ_t maps to learning about a hidden Markov chain from a sequence of normal signals. This problem is known as Kalman filtering (Kalman 1960) and implies the following property:

$$Var_t(\theta_t) = Var_{t-1}(\theta_t) - \frac{Cov_{t-1}(r_t, \theta_t)^2}{Var_{t-1}(r_t)} \tag{7.15}$$

which, joined to (7.14), simplifies to a single equation that determines the equilibrium

[24] Beyer et al. (2019) show that, in a finite horizon model as the number of periods increase, the coefficients of the price response and residual uncertainty will converge to the steady-state. Practically, one solves for the steady-state by dropping the time-dependence on all time-varying coefficients.

residual uncertainty (and, hence, the price response):

$$\sigma_\theta^2 = \rho^2\sigma_\theta^2 + \sigma_\varepsilon^2 - \frac{(\rho^2\sigma_\theta^2 + \sigma_\varepsilon^2)^2}{\rho^2\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\zeta^2}. \tag{7.16}$$

To estimate the model, it is useful to derive the model implications about the relationship between prices and reports. Unfortunately, the component $\mathbb{E}_{t-1}(r_t)$ in the price equation (7.12) is time-varying and cannot be recovered directly from the data. Hence, Beyer et al. (2019) solve for this term using the structure of the model:

$$\mathbb{E}_{t-1}(r_t) = \mathbb{E}_{t-1}(\frac{\beta\gamma}{c} + \rho\theta_{t-1} + \varepsilon_t + \zeta_t) = \frac{\beta\gamma}{c} + \rho\underbrace{\mathbb{E}_{t-1}(\theta_{t-1})}_{=p_{t-1}}. \tag{7.17}$$

Plugging this into the price equation, one arrives at

$$p_t = p_{t-1} + \gamma\beta(r_t - \frac{\beta\gamma}{c} - \frac{\rho}{\beta}p_{t-1}). \tag{7.18}$$

This model predicts a price that is a deterministic function of the report and the previous period's price. This is of course a prediction that would fail empirically as prices are noisy proxies of fundamentals; to be more explicit, there is a price discovery process that is not being captured in the structural model. Hence, to make the model amenable to empirical testing, one can add a noise term ν_t to the above equation and arrive at the following ARMAX model:

$$p_t = \beta_0 + \beta_1 p_{t-1} + \beta_2 r_t + \nu_t. \tag{7.19}$$

There are several ways of estimating this model. For example, one can estimate the values of $\beta_0 = -\frac{(\beta\gamma)^2}{c}, \beta_1 = 1 - \gamma\rho$ and $\beta_2 = \beta\gamma$. Note that this is a system of three equations in four unknowns (β, γ, ρ, c), so the linear model is not sufficient to identify the structural parameters. To address this problem, a common approach is to think about

conditional identification, writing first all parameters as a function of ρ, in the sense that one can infer $\gamma, \beta,$ and c from (7.19) if ρ is known. Only a single additional equation is needed to recover ρ and given that the pricing equation is primarily based on investors' Bayesian pricing, it is natural to seek identification of ρ from the investor's optimal reporting strategy. Rewriting the process for r_t in (7.13),

$$r_t = \frac{\beta\gamma}{c} + \rho r_{t-1} - \rho\zeta_{t-1} + \varepsilon_t + \zeta_t, \tag{7.20}$$

which is a simple ARMA process and yields estimates for ρ as well as the remaining noise terms σ^2_ζ and σ^2_ε.

7.3 Dynamic Models of Detection

This last application focuses on the use of value functions to estimate dynamic structural models with a full solution approach. The model developed here is based on a simplified version of Zakolyukina (2018).[25] The most important economic features of this model are that (1) manipulations are subject to random detection, (2) managers suffer a fixed and variable penalty when detected, (3) manipulations reverse in the next period, and (4) managers are forward-looking and consider the possibility of future detections. In addition, the approach assumes that concerns for manipulation are not a first-order determinant of compensation arrangements - i.e., manipulation is not so severe that it would substantially distort incentives. Hence, the problem can be formulated as maximizing the discounted payoff of manipulated earnings.

In the full model, variation in incentives may be due to many observable factors such as variation in wealth, risk-aversion or nearing retirement. To simplify the model to its

[25]The development of this model incorporates additional insights from dynamic misreporting with noise in the reporting process in Terry, Whited and Zakolyukina (2018) and McClure and Zakolyukina (2022). This type of model also has implications for productive efficiency - however, these require additional developments of neo-classical investment theory with incomplete information and is beyond the scope of this section: we refer to Terry (2015) and Terry et al. (2018) for recent treatments of these questions.

essential building blocks, we use below a random incentives formulation á la Fischer and Verrecchia (2000) in which all factors are represented as a random shock to incentives. Time is indexed by $t = 0, \ldots, \infty$ with the manager starting employment at date $t = 0$. The manager achieves a per-period payoff

$$\pi_t = (\mu_e + \epsilon_t)(b_t - b_{t-1}) - g_t \mathcal{C}_t(b_t), \tag{7.21}$$

where $\epsilon_t \sim N(0,1)$ is an i.i.d. shock to incentives, b_t is manipulation at date t which, after the reversal from the prior period manipulation, increases the manager's compensation by $(b_t - b_{t-1})$, g_t indicates a binary detection (to be defined later), and $\mathcal{C}_t(b_t)$ is a manipulation cost. A manager who has manipulated at time t or $t-1$, but has not been detected yet, has a probability $\gamma \in (0,1)$ of being detected ($g_t = 1$), and then bears a legal penalty

$$\mathcal{C}_t(b_t) = \kappa_1 + \kappa_2 b_t^2/2, \tag{7.22}$$

which includes a fixed component κ_1 and a variable component $\kappa_2 b_t^2$.

After observing ϵ_t, the manager chooses b_t to maximize

$$V_t \equiv \mathbb{E}_t(\sum_{t'=t}^{\infty} \beta^{t'-t} \pi_t | \epsilon_t), \tag{7.23}$$

where $\mathbb{E}_t(.)$ represents the expectation with respect to all information about past manipulations and can be summarized as a state b_t of undetected manipulation and the current incentive shock ϵ_t. Dropping time indices, we can then write the expected discounted payoff $V(b)$ to the manager given a manipulation b in the prior period before the incentive shock ϵ is observed (and let $'$ denote the current period):

$$V(b) = \mathbb{E}(\max_{b'} \underbrace{(\mu_e \epsilon)(b' - b) - \gamma d' \mathcal{C}(b') + \beta(1 - d' + \gamma d')V(0) + \beta d'(1-\gamma)V(b')}_{\equiv \Pi(b, b', \epsilon)}), \tag{7.24}$$

where $d' = 1_{b\neq 0|b'\neq 0}$ is equal to one when there is a non-zero manipulation in the prior period or the current period.

To numerically solve this model, the usual approach is to discretize the state space on a grid $B = (b_i)_{i=1}^n$ and $E = (\epsilon_i)_{i=1}^n$ and start from a guess about the value function. For example, a straightforward guess in this problem is that the manager does not manipulate, which implies $V^0(b) = V(0) = 0$. Then, value function iteration prescribes to repeatedly update these value functions using $V^i(.)$ and V_0^i in lieu of $V(b)$ and $V(0)$ in the right-hand side of (7.24). At each iteration, for each of the states b_i and ϵ_i, the optimal policy $b' \in B$ is calculated and an average expected payoff over the grid of shocks ϵ is calculated to recover an updated value function $V^{i+1}(b_i)$. Once all states $b_i \in B$ have been considered, $V^{i+1}(.)$ forms the new conjecture to use instead of $V^i(.)$, until a stopping criterion in the form of a norm $||V^{i+1} - V_i||$ sufficiently small indicates convergence to a solution of the dynamic program.

It is nevertheless rarely a good idea to jump directly into the estimation of a dynamic model without first forming intuition over the theoretical model. For this reason, we explore below some of the theoretical properties of the model. A common tool is to apply the envelope theorem on the Bellman equation to recover $V'(b)$. The envelope theorem states that the maximum in the right-hand side of (7.24) can be differentiated holding all other variables fixed. In the current problem, the right-hand side is differentiable if $b \neq 0$, which yields $V'(b) = -\mu_e$. Intuitively, the manager bears a cost (or benefit) of reversal equal to the expected incentive shock times per unit of manipulation.

This property also implies that the value function is linear in b (for any non-zero b) with a general form:

$$V(b) = 1_{b=0}V_0 + 1_{b\neq 0}V_1 - \mu_e b, \tag{7.25}$$

where $V_0 = V(0)$ and $V_1 = V(1) + \mu_e$ are two constants that do not depend on b.

The payoff of the manager when choosing the optimal manipulation strategy in (7.24)

can then be calculated as a function of these constants. A choice of $b' = 0$ yields a payoff

$$\Pi_0(b, \epsilon) \equiv -\gamma 1_{b\neq 0}\kappa_1 + \beta V_0 - b(\mu_e + \epsilon), \tag{7.26}$$

while a choice of $b' \neq 0$ yields a payoff $\Pi(b, b', \epsilon)$ concave and differentiable in b'. This latter expression can be maximized by taking a first-order condition in b' over the bracketed expression in (7.24) which, using that $V'(b') = -\mu_e$, implies

$$\mu_e + \epsilon - \beta\mu_e(1 - \gamma) - b'\gamma\kappa_2 = 0. \tag{7.27}$$

A greater probability of detection γ or manipulation cost κ_2 reduces manipulation. Solving (7.27) for b' and reinjecting into $\Pi(b, b', \epsilon)$ implies a payoff from a non-zero manipulation

$$\Pi_1(b, \epsilon) \equiv \frac{1}{2\gamma\kappa_2}\epsilon^2 + (\frac{(\beta(\gamma - 1) + 1)\mu_e}{\gamma\kappa_2} - b)\epsilon + \xi, \tag{7.28}$$

where ξ is a constant that does not depend on ϵ or b and whose expression is skipped to save space.[26]

The difference

$$\Delta(b, \epsilon) \equiv \Pi_1(b, \epsilon) - \Pi_0(b, \epsilon) = \frac{1}{2\gamma\kappa_2}\epsilon^2 + \frac{1 - (\beta(\gamma - 1))\mu_e}{\gamma\kappa_2}\epsilon + \xi + \gamma 1_{b\neq 0}\kappa_1 - \beta V_0 \tag{7.30}$$

captures the net benefit of non-zero manipulation and is illustrated in Figure 8.

The quadratic term in ϵ implies that the manager is more willing to manipulate for large incentive shocks because the benefit of greater manipulation offsets the penalties. Two additional properties of this model are essential to understanding the model quan-

[26] The constant ξ is similarly obtained from the first-order condition in (7.27) and is equal to

$$\xi = -b\mu_e - \gamma\kappa_1 + \frac{(\beta(\gamma - 1) + 1)^2\mu_e^2}{2\gamma\kappa_2} + \beta\gamma V_0 + V_1(\beta - \beta\gamma). \tag{7.29}$$

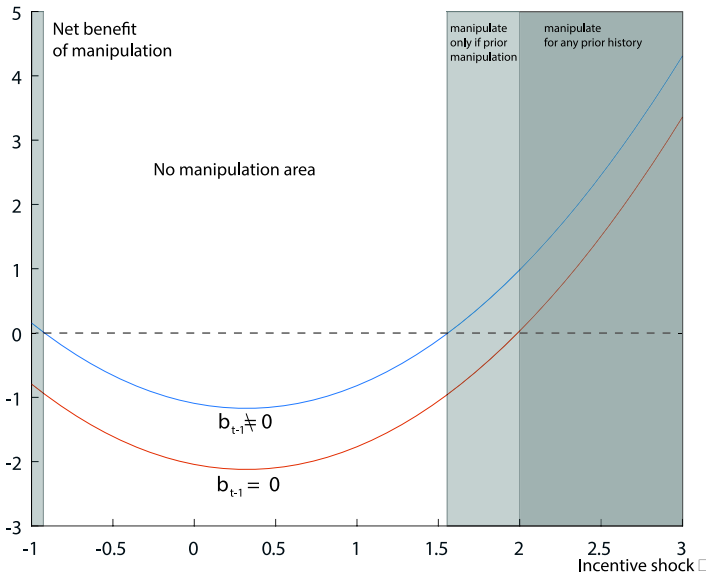


Figure 8: Net benefit of manipulation $\Delta(b, \epsilon)$

titative predictions about manipulation dynamics. A greater fixed penalty κ_1 increases the cost of starting of manipulation streak, and thus favors a choice of zero manipulation $\Pi_0(b, \epsilon)$. In addition, the term $1_{b\neq 0}\kappa_1$ in (7.30) implies that, when the manager starts manipulating and is already at risk of detection, manipulation becomes more likely in the next period. This parameter thus leads to strings of manipulation instead of uncorrelated single-period occurrences.

A useful exercise to understand the properties of the model is to simulate observables from the solution of the model. To conduct this simulation, we start from the state $b_0 = 0$, draw an incentive shock ϵ, apply the policy function b obtained from solving (7.24), randomly draw a detection and then update to the next state. Figure 9 plots a career of 30 time periods; he manager does not manipulate up to period 10, then over-reports for two periods only to revert to no-manipulation for three periods and this over-reporting spell is never caught. Then, the manager alternates between over-reporting, under-reporting and, ultimately, escapes detection by reverting to no manipulation at time 25. In the final periods, the manager starts of new spell of manipulation, and is finally detected, so that

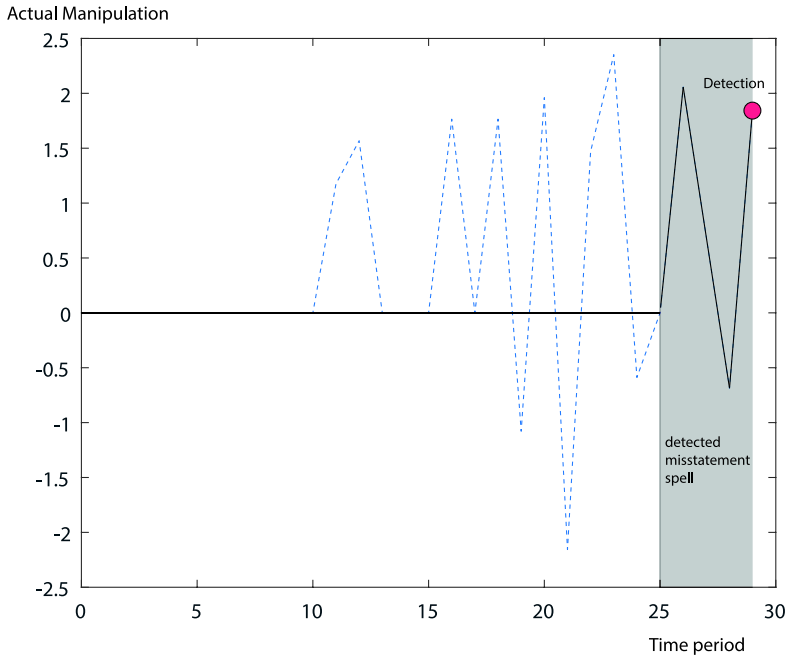


Figure 9: Simulated career

the spell of manipulation from period 26 to 29 is revealed. The observable is then the solid curve and will show, empirically, one restatement announcement and three manipulated periods, even though the manager manipulated 10 out of 30 periods.

Note that the theoretical analysis can also be used to simplify the numerical solution of the model, by observing that

$$V(b) = \mathbb{E}(\max(\Pi_0(b,\epsilon), \Pi_1(b,\epsilon)) - b\mu_e, \tag{7.31}$$

with a right-hand side that depends only on the two unknowns V_0 and V_1 rather than the entire function $V(b)$. Hence, one can also solve here a simplified Bellman equation by evaluating it at $b = 0, 1$ to obtain two equations in two unknowns.

We turn next to the identification of the model from observational data about detected misstatements. Identification refers to whether observable features of the data can be uniquely by one set of parameter values in the model. The ideal method to establish

identification is to prove analytically that there is an injective mapping from features of the data to parameters but, unfortunately, such proofs are typically infeasible in dynamic models because the models can only be solved numerically. Identification nevertheless poses certain challenges because a model that is poorly identified may pose certain problems such as failing to yield parameter estimates or standard errors or, more commonly, extreme sensitivity to numerical methods, samples or functional forms. Thus, researchers may wish to gain confidence that the estimation procedure suitably identifies the parameters of interest with feasible procedures that would detect some (but not all) identification problems.

The first approach involves simulating the model for many possible parameter values a-priori plausible and, using the simulated "laboratory" dataset, assess whether an econometrician unaware of these parameter values would be able to correctly recover these values. This procedure is thus a numerical analogue to proving identification theoretically but is only feasible for simple models in which estimation is fast and, therefore, a large enough subset of values can be explored.

To illustrate this approach, let us simulate a dataset with $50,000$ observations using a true parameter $\theta_0 = (.1, 3, 3)$. Note that, since identification is an asymptotic property and our objective here is not to assess the noise of an estimation procedure, we do not need to use here the same number of observations as the empirical dataset. The number of observations that can be used could be smaller than the data if the process of simulation is computationally intensive or larger if more precise estimates are needed to assess identification. One benefit of this method is that it is useable for any estimation method, even methods such as likelihood-based method that do not always have a straightforward intuition for the identification. But it is only practical for problems with a fast estimation procedure as re-estimating the model over a very large parameter space is infeasible for many problems.

We illustrate this method below in a simplified estimation problem, in which we would

Method	Estimate	Objective	Time
fminsearch, initialized at θ_0	(0.10, 3.22, 2.78)	.000	93s
fminsearch, initialized at $(.2, 2, 2)$	(0.08, 3.46, 4.69)	.018	105s
particleswarm, narrow search	(0.10, 3.15, 2.80)	.000	2,255s
particleswarm, wide search	(0.10, 3.40, 2.77)	.000	1,933s

Table 1: Convergence of search methods

like to estimate three parameters: the probability of detection γ as well a the two penalty parameters κ_1 and κ_2. All other parameters are assumed to be known. Note that, because identification is an asymptotic property and we are not interested in assessing sampling noise, it may be desirable to use a larger number of observations to be able to disentangle sampling noise from an identification problem.

Consider an estimation method in which we match the following vector of moments m_0: (a) the frequency of detected restatement years, (b) the probability of a detected misstatements conditional on a detected misstatement in the prior period, and (c) the mean restatement. The reason for choosing these moments will become clearer when applying the second approach but, for now, let us ask if the econometrician would be able to find out these three parameters.

The econometrician can use a simple method of moment procedure by calculating the sum of squared differences between moments in the data, versus moments implied by the model. In the current model, it is easiest to compute moments of the model by using the same simulation code, which yields $m(\theta)$. The estimation procedure is to find θ to match the moments:

$$\hat{\theta} \in \min_{\theta} (m_0 - m(\theta))'(m_0 - m(\theta)). \tag{7.32}$$

It is preferable to conduct this exercise using a local search algorithm with starting values points sufficiently far away from θ_0, or, when possible, use a global search algorithm with wide enough space.

To conduct this procedure, we solve (7.32) and recover $\hat{\theta}$ using Matlab. In Table 1, the results are given using multiple search algorithms: (i) a fast local search algorithm, fminsearch, starting at θ_0, (ii) the local search algorithm but starting from a more distant value $\theta = (.2, 2, 2)$, (iii) a global search algorithm, particleswarm, with a narrow search window from $(.05, 1, 1)$ to $(.2, 5, 5)$, (iv) the global search algorithm with a wide search window $(0, 0, 0)$ to $(1, 10, 10)$. Identification will fail if the minimum to the objective function is attained for values different from θ_0; in addition, the exercise can draw caution for some numerical methods if the search algorithms converges to a local minimum. In this analysis, constrained global and local search close enough to θ_0 recover the true parameter value. However, the analysis shows (a common feature in structural models) that global search without "help" on bounding the parameter search or local search from values that are too far from the data-generating process will fail to find an optimum. This problem is easily identified here by an objective function that is very different from zero; however, it is more difficult to diagnose if there are more moments than parameters. Because the objective function is rarely convex, inadequate local search algorithms (or, for that matter, global search algorithms with constraints that are too wide) can find incorrect parameter values.

This issue can imply that researcher priors about where the true parameters may lie, by explicitly or implicitly restricting the search space, may affect conclusions. This can be a benefit or cost, depending on one's point of view on the role of priors in scientific analysis. In all cases, making priors as explicit as possible (by making the search rule clear), rather than embedding them into a black box of numerical optimization, is always preferable.

A limitation of this approach is that it does not explain why the model is identified, and therefore does not guarantee that the estimation procedure is ideal for the model. For example, the model may be, in principle, identified from functional forms but, in practice, missing on other moments or other data features that pin down the economic mechanisms

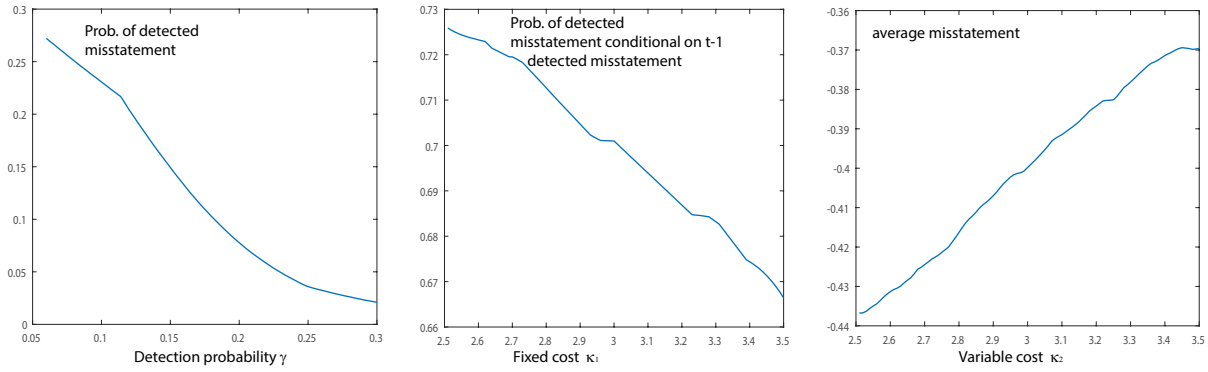


Figure 10: Moments in Simulated Data

of the model.

A second approach is to draw intuition about identification from the properties of the theoretical model. The detection parameter γ is the simplest and maps directly to a frequency of detections observed empirically. The second parameter κ_1 was shown in the theoretical analysis to lead to misstatement streaks, because a high κ_1 induces managers to be relatively more likely to manipulate again after an undetected manipulation. This parameter is therefore linked to the conditional probability of detected manipulation conditional on a detected manipulation in a prior periods. Lastly, the parameter κ_2 captures the disutility from increasing manipulation, and thus maps to the average manipulation. To assess these comparative statics quantitatively, it is useful to plot the moment as a function of their theoretically linked parameters. A non-monotonic or flat relation is a threat to identification. Figure 10 below suggests no obvious threat to identification.

Note that while both methods can only serve as necessary conditions to check for identification, they are, from an applied perspective, extremely powerful tools to detect failures of identification that may not be obvious in a complex models with many variables and moments. To give an example, consider the same problem but with a mean shock $\mu_e = 0$; in this case, it is easily verified analytically that that the manager should be equally likely to manipulate upwards or downwards, and, therefore, the mean manip-

ulation is zero regardless of κ_2. We are then left to identify three parameters with only two moments and identification is no longer possible. Even without a full understanding of the model, this problem is revealed by the two methods used above. First, a plot of the expected manipulation in the simulated data reveals that the moment is flat around zero (as expected theoretically), demonstrating that the moment does not adequately identify κ_2. Second, when running the global search algorithm in Table 1, the algorithms settles at $(.12, 9.22, .75)$, which conclusively suggests that the estimation fails.

8 Concluding Remarks

Accounting is concerned with the management of financial and managerial information. This is a rich and exciting area but it presents a notable challenge because information responds to economic problems unobserved to the researchers or market participants. Therefore, reduced-form facts that document how agents design and use information cannot be interpreted without a clear theory of the economics of the problem being solved, implying that, more than any other field, research in accounting that aims to go beyond fact documentation requires a solid understanding of theory.

The unprecedented growth of structural models in accounting are a revolution in the intellectual foundations of an area that has been historically descriptive (Gow et al. 2016). The methods present an opportunity to think not as story-tellers but as academics being precise about the economic forces connecting accounting and economic problems, in a manner that allows to understand the decision problems being solved by agent. As any subfield in fast development, there is still limited knowledge about what frameworks are likely to be better suited to explain the data, or even which empirical methods are best suited to theoretical problems relevant to accounting. In this respect, one should be open to explore multiple theories and methods, until a greater body of knowledge is being built. That will involve borrowing from areas in finance, marketing and economics, in which

structural has had an enormous influence, but will likely require to adapt the models and methods to make them amenable to new accounting questions.

In this survey, we have shown that the development of these methods has been quickly advancing in four principal areas of general audience interest: agency theory, disclosure theory, earnings management, and auditing. Yet, there are many other important accounting questions that do not easily fit within these areas, and the methods and examples developed here should serve as flexible tools and call for more research on many other central topics of accounting research.

Bibliography

Abadie, Alberto, and Guido W Imbens (2008) 'On the failure of the bootstrap for matching estimators.' *Econometrica* 76(6), 1537–1557

Andrews, Donald WK (2000) 'Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space.' *Econometrica* pp. 399–405

Arcidiacono, Peter, and Robert A Miller (2011) 'Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity.' *Econometrica* 79(6), 1823–1867

Ball, Ray, SP Kothari, and Valeri V Nikolaev (2013) 'Econometrics of the basu asymmetric timeliness coefficient and accounting conservatism.' *Journal of Accounting Research* 51(5), 1071–1097

Bao, Yang, Bin Ke, Bin Li, Y Julia Yu, and Jie Zhang (2020) 'Detecting accounting fraud in publicly traded us firms using a machine learning approach.' *Journal of Accounting Research* 58(1), 199–235

Basu, Sudipta (1997) 'The conservatism principle and the asymmetric timeliness of earnings.' *Journal of Accounting and Economics* 24(1), 3–37

Bazdresch, Santiago, R Jay Kahn, and Toni M Whited (2018) 'Estimating and testing dynamic corporate finance models.' *The Review of Financial Studies* 31(1), 322–361

Bellman, Richard (1966) 'Dynamic programming.' *Science* 153(3731), 34–37

Benmelech, Efraim, Eugene Kandel, and Pietro Veronesi (2010) 'Stock-based compensation and ceo (dis)incentives.' *Quarterly Journal of Economics* 125(4), 1769–1820

Bertomeu, Jeremy (2014) 'Incentive contracts, market risk and cost of capital.' *Contemporary Accounting Research, forth.*

Bertomeu, Jeremy, and Davide Cianciaruso (2016) 'Verifiable disclosure'

Bertomeu, Jeremy, and Edwige Cheynel (2015) 'Asset measurement in imperfect credit markets.' *Journal of Accounting Research* 53(5), 965–984

Bertomeu, Jeremy, and Iván Marinovic (2016) 'A theory of hard and soft information.' *The Accounting Review* 91(1), 1–20

Bertomeu, Jeremy, Anne Beyer, and Daniel J. Taylor (2016) 'From casual to causal inference in accounting research: The need for theoretical foundations.' *Foundations and Trends in Accounting* 10(2-4), 262–313

Bertomeu, Jeremy, Anne Beyer, and Daniel Taylor (2015) 'From casual to causal inference in accounting research: The need for theoretical foundations.' Technical Report, Foundations and Trends in Accounting, forthcoming.

Bertomeu, Jeremy, Anne Beyer, and Ronald A. Dye (2011) 'Capital structure, cost of capital, and voluntary disclosures.' *The Accounting Review* 86(3), 857–886

Bertomeu, Jeremy, Edward Li, Edwige Cheynel, and Ying Liang (2022a) ‘How uncertain is the market about managers’ reporting objectives? evidence from structural estimation’

Bertomeu, Jeremy, Edwige Cheynel, and Davide Cianciaruso (2021a) ‘Strategic withholding and imprecision in asset measurement.’ *Journal of Accounting Research* 59(5), 1523–1571

Bertomeu, Jeremy, Edwige Cheynel, Edward Xuejun Li, and Ying Liang (2021b) ‘How pervasive is earnings management? evidence from a structural model.’ *Management Science* 67(8), 5145–5162

Bertomeu, Jeremy, Edwige Cheynel, Eric Floyd, and Wenqiang Pan (2021c) ‘Using machine learning to detect misstatements.’ *Review of Accounting Studies* 26(2), 468–519

Bertomeu, Jeremy, Edwige Cheynel, Yifei Liao, and Mario Milone (2021d) ‘Using machine learning to measure conservatism’

Bertomeu, Jeremy, Iván Marinovic, Stephen J Terry, and Felipe Varas (2022b) ‘The dynamics of concealment.’ *Journal of Financial Economics* 143(1), 227–246

Bertomeu, Jeremy, John Harry Evans III, Mei Feng, and Ayung Tseng (2021e) ‘Tacit collusion and voluntary disclosure: Theory and evidence from the us automotive industry.’ *Management Science* 67(3), 1851–1875

Bertomeu, Jeremy, Paul Ma, and Iván Marinovic (2020) ‘How often do managers withhold information?’ *The Accounting Review* 95(4), 73–102

Beyer, Anne, and Ronald A. Dye (2012) ‘Reputation management and the disclosure of earnings forecasts.’ *Review of Accounting Studies* 17(4), 877–912

Beyer, Anne, Ilan Guttman, and Iván Marinovic (2019) 'Earnings management and earnings quality: Theory and evidence.' *The Accounting Review* 94(4), 77–101

Bird, Andrew, Stephen A Karolyi, and Thomas G Ruchti (2019) 'Understanding the numbers game.' *Journal of Accounting and Economics* 68(2-3), 101242

Borges, Jorge Luis (1998) 'On the exactitude of science. collected fictions.' *Translated by Andrew Hurley. New York: Penguin*

Breuer, Matthias (2021) 'How does financial-reporting regulation affect industry-wide resource allocation?' *Journal of Accounting Research* 59(1), 59–110

Breuer, Matthias, and David Windisch (2019a) 'Investment dynamics and earnings-return properties: A structural approach.' *Journal of Accounting Research* 57(3), 639–674

——— (2019b) 'Investment dynamics and earnings-return properties: A structural approach: Online appendix'

Breuer, Matthias, and Harm H Schütt (2019) 'Accounting for uncertainty: An application of bayesian methods to accruals models.' *Review of Accounting Studies, forth.*

Cartwright, Nancy (2007) *Hunting causes and using them: Approaches in philosophy and economics* (Cambridge University Press)

Caskey, Judson (2014) 'The pricing effects of securities class action lawsuits and litigation insurance.' *The Journal of Law, Economics, & Organization* 30(3), 493–532

Causholli, Monika, and W Robert Knechel (2012) 'An examination of the credence attributes of an audit.' *Accounting Horizons* 26(4), 631–656

Chemla, Gilles, and Christopher Hennessy (2021a) 'Equilibrium counterfactuals.' *International Economic Review* 62(2), 639–669

____ (2021b) ‘Signaling, instrumentation, and cfo decision-making.’ *Journal of Financial Economics*

Cheynel, E, and M Liu-Watts (2015) ‘Structural estimation of disclosure theory’

Cheynel, Edwige (2013) ‘A theory of voluntary disclosure and cost of capital.’ *Review of Accounting Studies* 18(4), 987–1020

Cheynel, Edwige, and Amir Ziv (2021) ‘On market concentration and disclosure.’ *Journal of Financial Reporting* 6(2), 1–18

Cheynel, Edwige, and Bertomeu Bertomeu (2020) ‘Accounting and the financial accelerator’

Cheynel, Edwige, and Frank Zhou (2020a) ‘Audit firm rotation and misstatements: A dynamic discrete choice approach.’ *Available at SSRN 3284807*

____ (2020b) ‘Audit firm rotation and misstatements: A dynamic discrete choice approach.’ *Available at SSRN 3284807*

Cheynel, Edwige, and M Liu-Watts (2020) ‘A simple structural estimator of disclosure costs.’ *Review of Accounting Studies* pp. 1–45

Cho, In-Koo, and David M. Kreps (1987) ‘Signaling games and stable equilibria.’ *Quarterly Journal of Economics* 102(2), 179–222

Choi, Jung Ho (2021) ‘Accrual accounting and resource allocation: A general equilibrium analysis.’ *Journal of Accounting Research* 59(4), 1179–1219

Core, John E, Wayne R Guay, and Robert E Verrecchia (2003) ‘Price versus non-price performance measures in optimal ceo compensation contracts.’ *The accounting review* 78(4), 957–981

David, Joel M, Hugo A Hopenhayn, and Venky Venkateswaran (2016) 'Information, misallocation, and aggregate productivity.' *The Quarterly Journal of Economics* 131(2), 943–1005

Dye, Ronald A. (1985) 'Disclosure of nonproprietary information.' *Journal of Accounting Research* 23(1), 123–145

—— (1988) 'Earnings management in an overlapping generations model.' *Journal of Accounting Research* 26(2), 195–235

Dye, Ronald A., and Sri S. Sridhar (2008) 'A positive theory of flexibility in accounting standards.' *Journal of Accounting and Economics* 46(2-3), 312–333

Dye, Ronald A., and Sri Sridhar (2004) 'Reliability-relevance trade-offs and the efficiency of aggregation.' *Journal of Accounting Research* 42(1), 51–88

Edmans, Alex, Vivian W Fang, and Katharina A Lewellen (2017) 'Equity vesting and investment.' *The Review of Financial Studies* 30(7), 2229–2271

Efron, B (1979) 'Bootstrap methods: Another look at the jackknife.' *The Annals of Statistics* 7(1), 1–26

Einhorn, Eti, and Amir Ziv (2008) 'Intertemporal dynamics of corporate voluntary disclosures.' *Journal of Accounting Research* 46(3), 567–589

—— (2012) 'Biased voluntary disclosure.' *Review of Accounting Studies* 17(2), 420–442

Fellingham, John C, D Paul Newman, and Yoon S Suh (1985) 'Contracts without memory in multiperiod agency models.' *Journal of Economic Theory* 37(2), 340–355

Fischer, Paul E., and Robert E. Verrecchia (2000) 'Reporting bias.' *The Accounting Review* 75(2), 229–245

Friedman, Jerome H (2002) ‘Stochastic gradient boosting.’ *Computational statistics & data analysis* 38(4), 367–378

Gayle, George-Levi, and Robert A. Miller (2005) ‘Has moral hazard become a more important factor in managerial compensation?’ Tepper School of Business Working Paper

Gayle, George-Levi, and Robert A Miller (2009) ‘Has moral hazard become a more important factor in managerial compensation?’ *The American Economic Review* 99(5), 1740–1769

——— (2015) ‘Identifying and testing models of managerial compensation.’ *The Review of Economic Studies* p. rdv004

Gayle, George-Levi, Chen Li, and Robert A Miller (2018) ‘How well does agency theory explain executive compensation?’

——— (2021) ‘Was sarbanes-oxley costly? evidence from optimal contracting on ceo compensation’

Gayle, George-Levi, Limor Golan, and Robert A Miller (2015) ‘Promotion, turnover, and compensation in the executive labor market.’ *Econometrica* 83(6), 2293–2369

Gerakos, Joseph, and Chad Syverson (2015) ‘Competition in the audit market: Policy implications.’ *Journal of Accounting Research* 53(4), 725–775

——— (2017) ‘Audit firms face downward-sloping demand curves and the audit market is far from perfectly competitive.’ *Review of Accounting Studies* 22(4), 1582–1594

Gopalan, Radhakrishnan, Todd Milbourn, Fenghua Song, and Anjan V Thakor (2014) ‘Duration of executive compensation.’ *The Journal of Finance* 69(6), 2777–2817

Gourieroux, Christian, Alain Monfort, and Eric Renault (1993) 'Indirect inference.' *Journal of applied econometrics* 8(S1), S85–S118

Gow, Ian D, David F Larcker, and Peter C Reiss (2016) 'Causal inference in accounting research.' *Journal of Accounting Research* 54(2), 477–523

Graham, John R., Campbell R. Harvey, and Shiva Rajgopal (2005) 'The economic implications of corporate financial reporting.' *Journal of Accounting and Economics* 40(1), 3–73

Grubb, Michael D (2011) 'Developing a reputation for reticence.' *Journal of Economics & Management Strategy* 20(1), 225–268

Guo, Qiang, Christopher Koch, and Aiyong Zhu (2017) 'Joint audit, audit market structure, and consumer surplus.' *Review of Accounting Studies* 22(4), 1595–1627

____ (2021) 'The value of auditor industry specialization: Evidence from a structural model.' *The Accounting Review*

Guttman, Ilan, Ohad Kadan, and Eugene Kandel (2006) 'A rational expectations theory of kinks in financial reporting.' *The Accounting Review* 81(4), 811–848

Hansen, Lars Peter (1982) 'Large sample properties of generalized method of moments estimators.' *Econometrica: Journal of the econometric society* pp. 1029–1054

Hansen, Lars Peter, and Kenneth J Singleton (1982) 'Generalized instrumental variables estimation of nonlinear rational expectations models.' *Econometrica: Journal of the Econometric Society* pp. 1269–1286

Hayashi, Fumio (2000) *Econometrics* (Princeton University Press)

Heckman, James J, and Bo E Honore (1990) 'The empirical content of the roy model.' *Econometrica: Journal of the Econometric Society* pp. 1121–1149

Hemmer, Thomas, Oliver Kim, and Robert E. Verrecchia (2000) 'Introducing convexity into optimal compensation contracts.' *Journal of Accounting and Economics* 28(3), 307–327

Hennessy, Christopher A, and Gilles Chemla (2022) 'Signaling, instrumentation, and cfo decision-making.' *Journal of Financial Economics* 144(3), 849–863

Hennessy, Christopher A, and Toni M Whited (2007) 'How costly is external financing? evidence from a structural estimation.' *The Journal of Finance* 62(4), 1705–1745

Holmström, Bengt (1979) 'Moral hazard and observability.' *Bell Journal of Economics* 10(1), 74–91

——— (1999) 'Managerial incentive problems: A dynamic perspective.' *Review of Economic Studies* 66(1), 169–182

Horowitz, Joel L (2001) 'The bootstrap.' In 'Handbook of econometrics,' vol. 5 (Elsevier) pp. 3159–3228

——— (2019) 'Bootstrap methods in econometrics.' *Annual Review of Economics* 11, 193–224

Hotz, V Joseph, and Robert A Miller (1993) 'Conditional choice probabilities and the estimation of dynamic models.' *The Review of Economic Studies* 60(3), 497–529

Hwang, Hyun, and Eunhee Kim (2019) 'A general equilibrium model of accounting standards.' In '103rd American Accounting Association Annual Meeting, AAA 2019'

Jensen, Michael C, and William H Meckling (1976) 'Theory of the firm: Managerial behavior, agency costs and ownership structure.' *Journal of financial economics* 3(4), 305–360

Jovanovic, Boyan (1982) 'Truthful disclosure of information.' *Bell Journal of Economics* 13(1), 36–44

Judd, Kenneth L (1998a) *Numerical methods in economics* (MIT press)

—— (1998b) *Numerical Methods in Economics* (MIT press, Cambridge)

Jung, Woon-Oh, and Young K. Kwon (1988) 'Disclosure when the market is unsure of information endowment of managers.' *Journal of Accounting Research* 26(1), 146–153

Kahn, Jay (2015) 'Influence functions for fun and profit.' *Ross School of Business, University of Michigan. Available from http://j-kahn. com/files/influencefunctions. pdf (version: July 10, 2015)*

Kahn, Robert, and Toni M Whited (2018) 'Identification is not causality, and vice versa.' *Review of Corporate Finance Studies* 7(1), 1–21

Kalman, Rudolph Emil (1960) 'A new approach to linear filtering and prediction problems'

Khan, Mozaffar, and Ross L Watts (2009) 'Estimation and empirical properties of a firm-year measure of accounting conservatism.' *Journal of accounting and Economics* 48(2-3), 132–150

Koopmans, Tjalling C (1975) 'Concepts of optimality and their uses. nobel memorial lecture.' In 'Three Essays on the State of Economic Science (New' Citeseer

Kydland, Finn E., and Edward C. Prescott (1982) 'Time to build and aggregate fluctuations.' *Econometrica* 50(6), 1345–1370

Lara, Juan Manuel Garc

MacKinnon, James G (2006) 'Bootstrap methods in econometrics.' *Economic Record* 82, S2–S18

Magnac, Thierry, and David Thesmar (2002) 'Identifying dynamic discrete decision processes.' *Econometrica* 70(2), 801–816

Mammen, Enno (1992) 'Bootstrap, wild bootstrap, and asymptotic normality.' *Probability Theory and Related Fields* 93(4), 439–455

——— (2012) *When does bootstrap work?: asymptotic results and simulations*, vol. 77 (Springer Science & Business Media)

Margiotta, Mary M., and Robert A. Miller (2000) 'Managerial compensation and the cost of moral hazard.' *International Economic Review* 41(3), 669–719

Marinovic, Ivan (2013) 'Internal control system, earnings quality, and the dynamics of financial reporting.' *The RAND Journal of Economics* 44(1), 145–167

Marinovic, Iván, and Felipe Varas (2016) 'No news is good news: Voluntary disclosure in the face of litigation.' *The RAND Journal of Economics* 47(4), 822–856

——— (2019) 'Ceo horizon, optimal pay duration, and the escalation of short-termism.' *The Journal of Finance* 74(4), 2011–2053

Marinovic, Iván, and Sri S Sridhar (2015) 'Discretionary disclosures using a certifier.' *Journal of Accounting and Economics* 59(1), 25–40

Marschak, Jacob (1974) 'Economic measurements for policy and prediction.' In 'Economic information, decision, and prediction' (Springer) pp. 293–322

McClure, Charles, and Anastasia A Zakolyukina (2022) 'Non-gaap reporting and investment.' *Chicago Booth Research Paper*

McFadden, Daniel (1973) ‘Conditional logit analysis of qualitative choice behavior.’ Technical Report

——— (1980) ‘Econometric models for probabilistic choice among products.’ *Journal of Business* pp. 13–29

Mehra, Rajnish, and Edward C. Prescott (1985) ‘The equity premium: A puzzle.’ *Journal of Monetary Economics* 15(2), 145–161

Mirrlees, James A (1999) ‘The theory of moral hazard and unobservable behaviour: Part i.’ *The Review of Economic Studies* 66(1), 3–21

Nikolaev, Valeri V (2019) ‘Identifying accounting quality.’ *Chicago Booth Research Paper*

Popper, Karl (2014) *Conjectures and refutations: The growth of scientific knowledge* (routledge)

Rust, John (1987) ‘Optimal replacement of gmc bus engines: An empirical model of harold zurcher.’ *Econometrica: Journal of the Econometric Society* pp. 999–1033

——— (1994) ‘Structural estimation of markov decision processes.’ *Handbook of econometrics* 4, 3081–3143

Stein, Jeremy C. (1989) ‘Efficient capital markets, inefficient firms: A model of myopic corporate behavior.’ *Quarterly Journal of Economics* 104(4), 655–669

Stokey, Nancy L, RE Lucas, and E Prescott (1989) ‘Recursive methods in dynamic economics.’ *Cambridge, MA: Harvard University*

Tauchen, George (1986) ‘Finite state markov-chain approximations to univariate and vector autoregressions.’ *Economics letters* 20(2), 177–181

Taylor, Lucian A (2010) ‘Why are ceos rarely fired? evidence from structural estimation.’ *The Journal of Finance* 65(6), 2051–2087

Terry, Stephen J (2015) ‘The macro impact of short-termism.’ *Discussion Papers* pp. 15–022

Terry, Stephen, Toni M Whited, and Anastasia A Zakolyukina (2018) ‘Information versus investment.’ *Available at SSRN 3073956*

Verrecchia, Robert E. (1983) ‘Discretionary disclosure.’ *Journal of Accounting and Economics* 5, 179–194

Viscusi, W Kip (1978) ‘A note on” lemons” markets with quality certification.’ *The Bell Journal of Economics* pp. 277–279

Vuong, Quang H (1989) ‘Likelihood ratio tests for model selection and non-nested hypotheses.’ *Econometrica: Journal of the Econometric Society* pp. 307–333

Whited, Toni (2021) ‘Mitsui center summer school in structural models’

Zakolyukina, Anastasia A (2018) ‘How common are intentional gaap violations? estimates from a dynamic model.’ *Journal of Accounting Research* 56(1), 5–44