Measuring the "Dark Matter" in Asset Pricing Models

Hui Chen Winston Wei Dou Leonid Kogan*

September 9, 2013

Abstract

Models of rational expectations endow agents with precise knowledge of the probability laws inside the models. This assumption becomes more tenuous when a model's performance is highly sensitive to parameters that are difficult to estimate using available data, i.e., when a model relies on "dark matter." In this paper, we propose new measures of model fragility by quantifying the informational burden that a rational expectations model places on the agents. By measuring the informativeness of the cross-equation restrictions implied by a model, our measures can systematically detect the direction in the parameter space in which the model's performance is the most fragile. Our methodology can be used to conduct sensitivity analysis on quantitative models. It helps identify situations where parameter or model uncertainty cannot be ignored. It can also be used to evaluate competing classes of models that try to explain the same set of empirical phenomena from the perspective of the robustness of their implications.

*Chen: MIT Sloan and NBER (huichen@mit.edu). Dou: MIT Sloan (wdou@mit.edu). Kogan: MIT Sloan and NBER (lkogan@mit.edu). We thank Murray Carlson, John Cochrane, Ian Drew-Becker, Lars Hansen, Martin Schneider, Laura Veldkemp, Fernando Zapatero, Stan Zin and seminar participants at MIT Sloan, INSEAD, ITAM, the Chicago Initiative in Theory and Empirics Conference, the UBC Winter Finance Conference, the Macrofinance Workshop, WFA, and the NBER Working Group on the Capital Markets and the Economy for comments.

1 Introduction

The assumption of rational expectations (RE) is a predominant and powerful technique in quantitative economic analysis. It ties down the beliefs of economic agents by endowing them with precise knowledge of the probability law implied by an economic model. This assumption is usually justified as a limit resulting from a sufficiently long history of learning from a wealth of data, which allows the model builders to presume RE as an approximation of the true belief formation process (see Hansen (2007)).[1] While the intention of the RE assumption is to discipline agents' beliefs, its use in practice sometimes implies the opposite. For example, if the model output is sensitive to small changes in a parameter that is weakly identified in the data, assuming precise knowledge of the parameter essentially gives the modeler an additional degree of freedom. In this paper we attempt to quantify the degree of fragility of the RE assumption by measuring the informational burden it places on the economic agents.

To fix the ideas, consider a representative-agent model designed to explain the observed dynamics of certain asset prices. As a part of the explanation, the model appeals to the dynamics of fundamental data described by a known probability law parameterized by θ. Under the rational expectations assumption, the representative agent knows the true value of the parameter vector, θ_0. The RE equilibrium can be viewed as a tractable approximation to a more general economy, in which the representative agent maintains nontrivial uncertainty about the true parameter values based on all the information available and continues to learn from new data.

A necessary condition for the RE assumption to be a good approximation in this context is that the main implications of the RE model for the joint dynamics of prices and fundamentals should not be sensitive to the exact parameter values the agent contemplates as similarly likely to θ_0, given the remaining uncertainty she faces. Without this condition, we cannot claim that the uncertainty the agent maintains about θ_0 has no impact on prices. In other words, a RE model with parameters θ_0 is

[1] See Blume and Easley (2010) for a survey of the literature on convergence to rational expectations equilibria in models with various belief formation processes.

not a robust explanation of an empirical phenomenon if a RE model with parameters θ', which the agent considers just as plausible based on the information available, produces drastically different model predictions.

The above criterion connects the validity of the RE assumption to the subjective beliefs of the agent. Applying the criterion requires a plausible specification of the agent's beliefs regarding θ. A natural benchmark for such beliefs is the belief an agent would form based on available historical data on fundamentals. Under this benchmark, the economic agent is placed on a roughly equal footing with the econometrician in terms of their information sets. We label a RE model that fails to satisfy the above robustness requirement under plausible subjective beliefs as fragile.

Our robustness requirement is relatively weak. Passing the above requirement is necessary but not sufficient for justifying the RE approximation. In particular, even if the model produces similar implications for prices under all of the *relatively likely* parameter values, ignoring the uncertainty faced by the representative agent may still distort the key implications of the model. The reason is that in a model in which the agent entertains uncertainty about the parameter values, prices depend on the entire subjective distribution over the parameter values, hence low-probability parameter configurations may have a disproportionately large effect on prices.

It is convenient to re-state our robustness criterion in a slightly different form. After learning about the parameters from fundamental data, the agent will consider a subset of the parameter values as more likely. If the model produces similar implications for prices under all of these likely parameter values, an econometrician with access to the same set of fundamental data should not derive significant incremental information about the parameters by imposing cross-equation restrictions on fundamental data and prices. In other words, the appearance of cross-equation restrictions in a RE model being too informative about certain parameters relative to the information derived from fundamental data is a flip side of the lack of model robustness: those tightly restricted yet hard-to-measure inputs form the "dark matter" that is critical for the model's predictions.

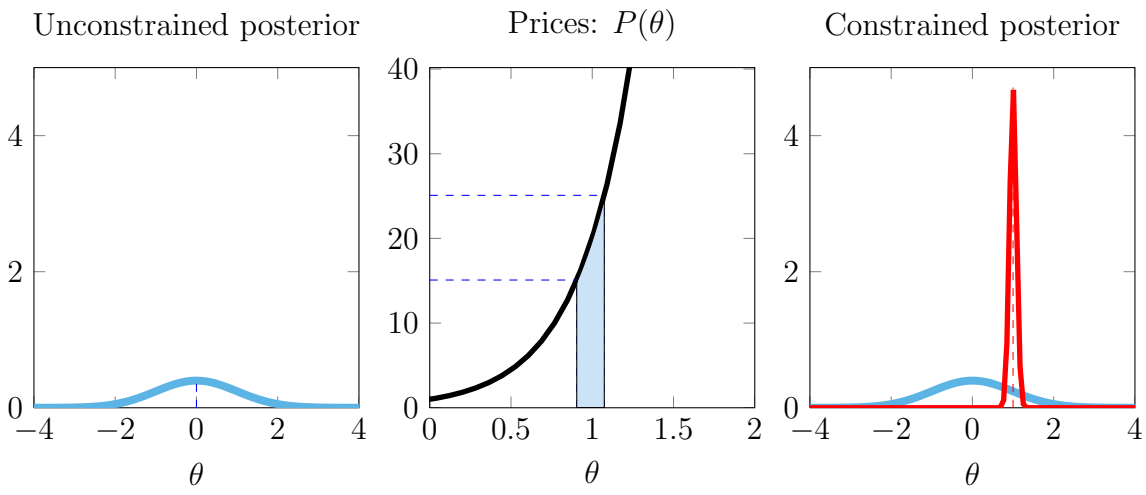


Figure 1: An example of "informative" cross-equation restrictions.

Figure 1 captures the essence of the above discussion. An econometrician observes fundamental data and forms a posterior belief about a parameter θ, which is represented by the distribution in the left panel. Suppose that the prices P implied by a RE model are sensitive to θ, i.e., the derivative $\partial P/\partial\theta$ is large enough that the model-implied prices $P(\theta)$ are statistically consistent with the price data only over a very small set of values of θ (see the middle panel). In this case, by imposing the relationship between prices and θ implied by the model, or, in other words, a cross-equation restriction, the econometrician obtains a constrained posterior for θ (see the right panel) that is much more concentrated than the distribution she was able to establish using only the fundamental data. This discrepancy shows that the model's pricing restriction is highly informative about the parameter θ. It also suggests that because prices are so sensitive to θ, the RE model that ignores any parameter uncertainty is not a robust approximation of a model in which the agent maintains realistic uncertainty about θ.

Ignoring the considerations of robustness when applying the RE framework can lead to misleading conclusions. The RE model appears successful since it "explains" asset prices with a carefully calibrated value of θ. As the flat distribution in the left panel of Figure 1 indicates, based on the fundamental data alone, there is a wide range of parameter values we cannot reject using standard statistical methods such as

likelihood ratio tests, and that includes the value picked by the RE model. Yet, our discussion highlights the lack of robustness of such an explanation – model implications change drastically when we change θ by a "small amount", measured with respect to the information available about θ from the fundamental data. Thus, the success of the model crucially depends on its "dark matter" inputs — the particular belief it bestows on the agent.

To operationalize the idea, we develop two measures to quantify the informativeness of the cross-equation restrictions. The first measure is the asymptotic information ratio, which applies in large sample. This measure is based on the comparison between the Fisher information matrix implied by the unconstrained and the constrained likelihood functions (the constraint being the cross-equation restrictions). Intuitively, this information ratio shows how much more difficult it is to estimate various smooth functions of model parameters in large samples without imposing cross-equation restrictions than with the restrictions, which is based on the amount of extra data needed (on average) to raise the precision of the unconstrained estimator to the same level as the constrained estimator. This measure can also be interpreted in terms of asymptotic detection rates – it quantifies asymptotically the gain in the econometrician's ability to distinguish statistically various parameter configurations that results from imposing cross-equation restrictions.

Our second measure is the finite-sample information ratio. In finite sample, the informativeness of cross-equation restrictions is reflected in their impact on the constrained posterior distribution of model parameters. We quantify the discrepancy between the unconstrained and constrained posterior distributions of model parameters using the relative entropy of the two distributions. Then, we provide a sample-size interpretation of our measure by finding the average amount of extra fundamental data needed to generate as large a shift – measured by relative entropy – in the posterior parameter distribution implied by the unconstrained model. Finally, we establish asymptotic equivalence between our two measures in a sense we make precise below.

Our results can be interpreted as a rigorous extension of the common practice

of using sensitivity analysis to assess the robustness of model implications. The key implications of a model are considered robust if they are not excessively sensitive to small perturbations in model parameters. However, such practice is ad hoc both in terms of how to determine the relevant perturbation magnitude and how to define "excessive sensitivity." Moreover, it is difficult to generalize the traditional sensitivity analysis to multivariate settings. Model fragility is generally not fully revealed by perturbing parameters one at a time – one must contemplate all possible multivariate perturbations, making the ad hoc approach essentially infeasible for high-dimensional problems. Our methodology overcomes these difficulties. In fact, as one of the outputs of the asymptotic information ratio calculation, we identify the direction in the parameter space in which the model is most fragile.

Another approach to judging robustness of a model is to test it using additional data. If a model is misspecified, further testable restrictions may reveal that. This approach has limitations as a general tool for our purposes. In many cases it takes subjective judgement to determine what predictions of the model should be tested to provide meaningful evidence of misspecification, since any model is only an approximate description of reality. More importantly, a model that passes certain additional tests may still be fragile in a sense of placing too much informational burden on the agents.

As an illustration, we use the fragility measures to analyze the robustness of a class of disaster risk models. In these models, the parameters that are difficult to estimate from the data are those describing the likelihood and the magnitude of the disasters. For any given value of the coefficient of relative risk aversion γ, we compute the asymptotic information ratio for all "acceptable" calibrations, i.e., parameter values that cannot be rejected by the consumption data based on a likelihood ratio test. With $\gamma = 3$, the asymptotic information ratio ranges from 24.5 to 37,000. This means that the economic agent within a model would require a data sample of 24 to 37,000 times the length of the available historical sample in order to match the information content of the pricing restrictions implied by the model. The information ratio drops with higher risk aversion. When $\gamma = 24$, the lowest asymptotic information ratio is 1.8.

We obtain similar results using the finite-sample information ratio. Thus, according to our measures, model specifications with relatively high risk aversion coefficients are much less fragile than those with low γ.

We then consider the case where the risk aversion coefficient γ is estimated jointly with the rest of the parameters. We decompose the information provided by the cross-equation restrictions into one part about γ and the other about the disaster parameters. With an uninformative prior on γ, we find that asset prices are informative mostly about γ, whereas the information ratio for the disaster probability and size is close to 1. In contrast, a prior that strongly favors low values of γ leads to large information ratios on the disaster parameters. These results further highlight the relatively high fragility of models that attach high ex ante likelihood to smaller values of the risk aversion parameter.

There are a few alternative interpretations of highly informative cross-equation restrictions. First, agents might indeed have stronger beliefs about certain parameters than what can be justified by the data available to the econometrician. These beliefs may not be anchored to the data (e.g., driven by investor sentiment), in which case our information criteria no longer apply. However, if strong beliefs are to be justified within the rational expectations framework, then understanding their sources (such as extra data or aggregation of private information) should be viewed as a key requirement for declaring a model as an explanation of the data. For example, in the context of rare disasters, Barro (2006), Barro and Ursua (2011), and Nakamura, Steinsson, Barro, and Ursa (2012) explore international macroeconomic data, which provide further information about disaster risk.

Second, informative cross-equation restrictions may imply that the canonical RE assumption, under which agents have precise ex ante knowledge of the model parameters, is not an acceptable approximation to the belief-formation process. Prices may appear informative because we have ignored important uncertainty the agent faces when we adopt the RE assumption. Instead, the model should explicitly describe the evolution of agents' beliefs, e.g., through Bayesian learning. Our methodology

extends naturally to such settings and allows us to draw inference about the fragility of the model's predictions with respect to the assumed prior beliefs of the agents.

Third, informative cross-equation restrictions could imply model misspecification beyond parameter uncertainty. For example, the disaster risk model under consideration may be omitting some economically significant sources of risk – incorporating them into the model could make the model less fragile, meaning that consumption disasters need not be as large and rare (not so dark) as implied by the original specification.

Hansen (2007) discusses extensively concerns about the informational burden that rational expectations models place on the agents, which is one of the key motivations for research in Bayesian learning, model ambiguity, and robustness.[2] In particular, the literature on robustness in macroeconomic models (see Hansen and Sargent, 2008; Epstein and Schneider, 2010, for a survey of this literature) recognizes that the traditional assumption of rational expectations is not reasonable in certain contexts. This literature explicitly generalizes such models to incorporate robustness considerations into agents' decision problems. Our approach is complementary to this line of research in that we propose a general methodology for measuring and detecting the fragility of rational expectations models, thus identifying situations in which parameter uncertainty and robustness could be particularly important, but we do not take a stand on how fragile models need to be modified.

Our work is related to the broader body of work on the effects of parameter uncertainty in economic models. In the particular context of rare disaster models, Weitzman (2007) argues that instead of the disaster risk being represented by a rare event with a small, precisely known probability, such risk arises naturally because of the agents' imperfect knowledge of the tail distribution of consumption growth. Wachter (2008) and Collin-Dufresne, Johannes, and Lochstoer (2013) analyze the effect of learning in such models. They find that time-variation in estimated disaster probabilities has a first-order effect on equilibrium dynamics.

[2]See Gilboa and Schmeidler (1989), Epstein and Schneider (2003), Hansen and Sargent (2001, 2008), and Klibanoff, Marinacci, and Mukerji (2005), among others.

Our work is connected to the literature in rational expectations econometrics, where the cross-equation restrictions have been used extensively to gain efficiency in estimating the structural parameters.[3] For the econometrician, the danger of imposing RE in the presence of "dark matter" is that it can lead to unjustified tight confidence intervals for the constrained estimator that will likely fail to contain the true parameter value. More broadly, the common practice of post-selection inference can become quite misleading in the presence of "dark matter".[4] Our information measures can be used to guard against selecting fragile models in such practice.

Given the basic nature of the problem, our methodology has a broad range of applications. In addition to its applications in the macro-finance area, it should be useful in evaluating and estimating structural models in many other areas of economics.

2 Illustration with a Disaster Risk Model

Before introducing the fragility measures, we use a simple example to illustrate scenarios in which the informativeness of the cross-equation restrictions in a rational expectations model serves as a signal for model fragility.

2.1 Model setup

We consider an endowment economy that is exposed to the risks of rare economic disasters. The setting is similar to Barro (2006). The log growth rate of aggregate consumption g_t and the excess log return of the market portfolio r_t follow the process

$$\begin{pmatrix} g_t \\ r_t \end{pmatrix} = (1 - z_t)u_t - z_t \begin{pmatrix} v_t \\ bv_t + \epsilon_t \end{pmatrix}, \tag{1}$$

[3]For classic examples, see Saracoglu and Sargent (1978), Hansen and Sargent (1980), Campbell and Shiller (1988), among others, and textbook treatments by Lucas and Sargent (1981), Hansen and Sargent (1991).

[4]Berk, Brown, Buja, Zhang, and Zhao (2012) argue that the selected model based on data-driven methods is itself stochastic, which should be incorporated into inference of parameters.

where z_t is i.i.d. Bernoulli and takes the value of 1 and 0 with probability p and $1-p$. Outside of disasters ($z_t = 0$), g_t and r_t are jointly normal with mean (μ, η_t), where μ and η_t are the expected consumption growth and conditional expected excess return in a non-disaster state (or conditional equity premium), respectively. Their covariance in the non-disaster state is

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}. \tag{2}$$

In the case of a disaster ($z_t = 1$), the log of the decline in consumption v_t follows a truncated exponential distribution, $v_t \sim 1_{\{v \geq \underline{v}\}} \lambda e^{-\lambda(v-\underline{v})}$, with the lower bound for disaster size equal to $\underline{v}$. Thus, conditional on a disaster, the average disaster size is $\underline{v} + 1/\lambda$. The log excess return in a disaster is linked to the decline in consumption with a leverage factor b. In addition, we add an independent shock $\epsilon_t \sim N(0, \nu^2)$ to r_t so that the excess return in a disaster does not have to be perfectly correlated with aggregate consumption.

The representative agent knows the value of all the parameters in this model except for p, and he observes z_t.[5] We denote the true value for p to be p_0. We consider two different specifications of the agent's beliefs about p. First, under rational learning, he starts with an uninformative prior about p at $t = 0$, which follows a Beta distribution,

$$\pi_0(p) = \mathbf{Beta}(\alpha_0, \beta_0) = p^{\alpha_0 - 1}(1-p)^{\beta_0 - 1}, \tag{3}$$

and he updates the belief about p over time based on the observed history of z_t. Then, at time t, his posterior belief about p will still follow a Beta distribution,

$$\pi_t(p) = \mathbf{Beta}(\alpha_t, \beta_t), \tag{4}$$

[5] This assumption helps simplify the analysis of the learning problem. In particular, the history of z_t is sufficient for Bayesian updating on p. Collin-Dufresne, Johannes, and Lochstoer (2013) analyze a disaster risk model where agents learn about multiple parameters simultaneously.

with

$$\alpha_t = \alpha_{t-1} + z_t, \quad \beta_t = \beta_{t-1} + (1 - z_t). \tag{5}$$

The posterior belief implies that the conditional expectation of p is $\mathbb{E}_t^a[p] = \alpha_t/(\alpha_t+\beta_t)$, where $\mathbb{E}^a$ denotes that the expectation is taken under the agent's information set.

Second, we also consider the case where the agent has precise but irrational belief about the disaster probability. Specifically, $\mathbb{E}_t^a[p] = p_1$, where p_1 is different from p_0.

The representative agent has time-additive isoelastic utility: $u(c) = c^{1-\gamma}/(1-\gamma)$, where $\gamma > 0$ is the coefficient of relative risk aversion. The consumption Euler equation for log excess returns is:

$$1 = \mathbb{E}_t^a \left[\frac{m_{t+1}}{\mathbb{E}_t^a[m_{t+1}]} e^{r_{t+1}} \right] = \mathbb{E}_t^a \left[\frac{(C_{t+1}/C_t)^{-\gamma}}{\mathbb{E}_t^a[(C_{t+1}/C_t)^{-\gamma}]} e^{r_{t+1}} \right]. \tag{6}$$

The Euler equation implies that the conditional equity premium is approximately[6]

$$\eta_t \approx \gamma\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma\mu - \frac{\gamma^2\sigma^2}{2}} \lambda \left(\frac{e^{\gamma \underline{v}}}{\lambda - \gamma} - e^{\frac{1}{2}\nu^2} \frac{e^{(\gamma-b)\underline{v}}}{\lambda + b - \gamma} \right) \frac{\mathbb{E}_t^a[p]}{1 - \mathbb{E}_t^a[p]}. \tag{7}$$

The first two terms on the right hand side give the risk premium for the exposure of the market portfolio to Gaussian shocks in consumption growth (with convexity adjustment). The third term gives the disaster risk premium. We need $\lambda > \gamma$ for the risk premium to be finite, which sets an upper bound for the average disaster size and dictates how fat the tail of the disaster size distribution can be.

The fact that the risk premium becomes unbounded as λ approaches γ is a key feature of this model. In particular, when the (perceived) probability of disaster is small, we can still generate a high risk premium by increasing the average disaster size (reducing λ), although a smaller λ will also make the risk premium more sensitive to changes in $\mathbb{E}_t^a[p]$. This sensitivity is crucial in our discussion of model fragility.

[6]The approximation we make here is $e^{\eta_t + \frac{\tau^2}{2} - \gamma\rho\sigma\tau} \approx 1 + \eta_t + \frac{\tau^2}{2} - \gamma\rho\sigma\tau$, which works for the parameter values we consider.

2.2 Learning, irrational beliefs, and rational expectation

We use the disaster risk model above as a laboratory to analyze whether the RE assumption provides a good approximation of the true model. Under the RE assumption, the agent knows the true value of p, which is denoted as p_0. Then, the equity premium outside of disasters is constant and is obtained by setting $\mathbb{E}_t^a[p] = p_0$ in (7):

$$\eta \approx \gamma\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma\mu - \frac{\gamma^2\sigma^2}{2}}\lambda\left(\frac{e^{\gamma\underline{v}}}{\lambda-\gamma} - e^{\frac{1}{2}\nu^2}\frac{e^{(\gamma-b)\underline{v}}}{\lambda+b-\gamma}\right)\frac{p_0}{1-p_0}. \tag{8}$$

We consider two related tests from the perspective of the econometrician. First, we examine how reliable statistical inference about the disaster probability p is under the RE assumption. Second, we examine how often the econometrician can reject the RE model by performing a standard specification test.

Given the data on consumption and returns, the econometrician can estimate p under both the unconstrained model and the RE model. According to the unconstrained model (1), the only relevant information for the estimation of p is the history of z_t. The maximum likelihood estimator (MLE) for p based on n observations $(z_1, \cdots, z_n)$ is asymptotically normal,

$$\widehat{p} \xrightarrow{D} N\left(p_0, \frac{p_0\left(1-p_0\right)}{n}\right). \tag{9}$$

According to the RE model (with the cross-equation restriction (8) imposed in the model (1)), the econometrician estimates p using both the macro data (g_t, z_t) and the returns r_t. In this case, the MLE of p has the following asymptotic distribution:

$$\widehat{p}^c \xrightarrow{D} N\left(p_0, \frac{1}{n}\frac{1}{\frac{1}{p_0(1-p_0)} + (1-p_0)\frac{\eta'(p_0)^2}{(1-\rho^2)\tau^2}}\right), \tag{10}$$

where

$$\eta'(p) \equiv e^{\gamma\mu - \frac{\gamma^2\sigma^2}{2}}\lambda\left(\frac{e^{\gamma\underline{v}}}{\lambda-\gamma} - e^{\frac{1}{2}\nu^2}\frac{e^{(\gamma-b)\underline{v}}}{\lambda+b-\gamma}\right)\frac{1}{(1-p)^2}$$

is the sensitivity of the equity premium with respect to the disaster probability p.

To measure how much extra precision is gained in the estimation of p by imposing the cross-equation restriction, we compute the ratio of the asymptotic variance of the unconstrained estimator (9) to that of the constrained estimator (10),

$$\varrho = 1 + \frac{\eta'(p)^2}{(1-\rho^2)\tau^2} p(1-p)^2 \,. \tag{11}$$

Equation (11) shows that, holding p fixed, the cross-equation restriction is highly informative about p when the model-implied equity premium is highly sensitive to the disaster probability ($\eta'(p)$ is large) and the noise term in returns has low variance ($(1-\rho^2)\tau^2$ is small).

To conduct the experiment, we simulate data (g_t, z_t, r_t) from 0 to $T_0 + T$ according to a specific model of beliefs (the true model). The econometrician only observes the data from T_0 to $T_0 + T$, whereas the agent observes the entire history of data. Having $T_0 > 0$ captures the possibility that the agent has more information than the econometrician. We study three alternative specifications of the agent's beliefs:

Model 1: Rational learning with large T_0. In this case, the agent has significant amount of information about p by time T_0, which is reflected in an informative prior $\pi_{T_0}(p)$.

Model 2: Rational learning with small T_0. In this case, the agent maintains significant uncertainty about p by T_0.

Model 3: Irrational beliefs. In this case, the agent believes in a wrong value for the disaster probability, $p_1 \neq p$, and does not update his belief.

We set the sample size for the econometrician to $T = 100$ to mimic the sample size of the U.S. annual consumption data. The risk aversion parameter is fixed at $\gamma = 4$, and the leverage parameter is $b = 3$. The lower bound of log disaster size $\underline{v}$ is 7%, and we set $z_t = 1$ in years when consumption growth is lower than -7%. The remaining parameters are calibrated to the aggregate consumption and market return

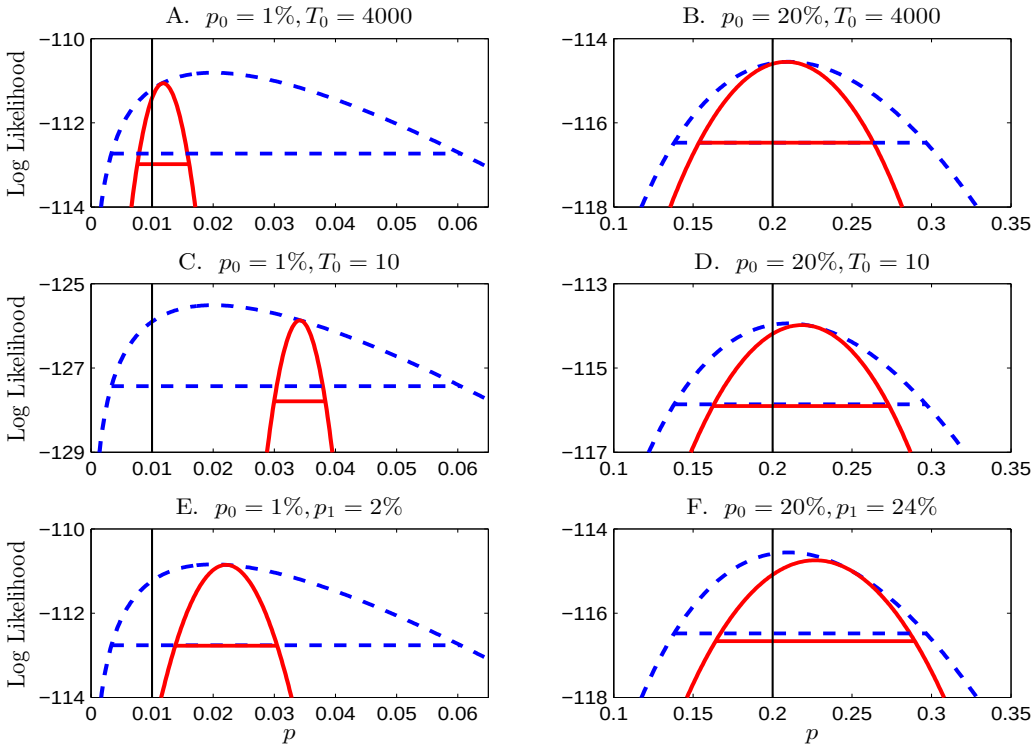


Figure 2: **Unconstrained vs. constrained likelihood.** The blue dash lines are the unconstrained log likelihood functions, while the red solid lines are the constrained log likelihood functions. The horizonal lines on the log likelihood functions mark the 95% confidence regions for p. In Panels A and B, the agent has observed 4000 years of extra data by T_0. In Panels C and D, the agent has only observed 10 years of extra data by T_0. In Panel E and F, the agent has irrational belief $p_1 \neq p_0$.

data: $\mu = 1.87\%$, $\sigma = 1.95\%$, $\eta = 5.89\%$, $\tau = 19.1\%$, $\rho = 59.4\%$, $\nu = 34.9\%$. For illustration, we consider two values of disaster probability p, 1% and 20%, and then solve for the disaster size parameter λ to make the model-implied equity premium in (8) equal to the sample mean of 5.89%. The corresponding values are $\lambda = 4.82$ and $\lambda = 63.06$.

In Figure 2, we plot the log likelihood function of p in the unconstrained model and the constrained RE model for a particular set of simulated data that is representative. The three rows correspond to Model 1 through Model 3 listed above, and the two columns correspond to $p_0 = 1\%$ and $p_0 = 20\%$.

The first thing to notice is that, under the same true disaster probability p_0, the

unconstrained likelihood functions are identical (except for a level difference) across the three models of beliefs. The 95% confidence intervals for p are marked by the horizontal dash lines. In the cases where $p_0 = 1\%$, the 95% confidence interval spans from 0.3% to 6.1%. When $p_0 = 20\%$, the confidence interval spans from 13.8% to 29.7%. Thus, while the confidence intervals indeed cover the true values, there is significant uncertainty remaining for the econometrician when she estimates p using the unconstrained model.

When the true value p_0 is small, the constrained likelihood function is much more concentrated than the unconstrained likelihood function. In contrast, when p_0 is large, the constrained likelihood function is only slightly more concentrated than the unconstrained likelihood function. The ratio of asymptotic variances in (11) explains this result. When p_0 is small, the model requires a large average disaster size (small λ) to match the observed equity premium, which makes η highly sensitive to changes in p. As p_0 increases, the average disaster size becomes smaller, and so does $\eta'(p)$. Based on our calibration, $\varrho = 20.7$ when $p = p_0 = 1\%$, and $\varrho = 2.0$ when $p = p_0 = 20\%$.

While the cross-equation restriction appears to provide substantial information about p in all three models of beliefs when p_0 is small, the added information is not always valid. In Model 1 (Panels A and B), the agent's prior at time T_0 is highly concentrated at the true value of p due to learning from a long history of data. Thus, returns generated by the model with learning are very close to those generated by the RE model. As a result, the cross-equation restriction from the RE model is approximately valid. This is reflected in the fact that the 95% confidence interval based on the constrained model covers the true value of p. The fact that the confidence interval is significantly tighter for small p_0 in Panel A shows that, under the condition that the RE model is a good approximation of the true model, having a fragile model (in the sense that asset pricing moments are very sensitive to small changes in the parameters) is helpful for inference.

Next, in Model 2 (Panels C and D), the agent faces significant uncertainty about p at T_0. His posterior beliefs about p change significantly with the arrival of new

observations, which is reflected in the observed series of asset returns. For this reason, the RE assumption is not a good approximation to the true model. In this case, imposing the cross-equation restriction in a fragile model (when p_0 is small) gives the econometrician false confidence in her estimate of p, which may be far away from the true value. Finally, in Model 3 (Panels E and F), the agent irrationally believes in a wrong value for disaster probability. In this case, the cross-equation restriction from the RE model is invalid by construction. When p_0 is small, the 95% confidence interval for the constrained model is again misleadingly tight and fails to cover the true parameter value.

The RE model is misspecified relative to each of the three models of beliefs. However, the misspecification may be difficult to detect with limited data. To quantify the ability of an econometrician to detect model misspecification, we simulate multiple samples of data (with sample size of 100) based on each model of beliefs, run the likelihood ratio (LR) tests of the RE restrictions, and report the rejection ratio — the fraction of the LR tests that reject the RE assumption. To perform the LR test, we specify the RE model as a special case of a general class of models. Consider the following model of returns that nests the model with rational learning, the model with irrational beliefs, and the model with RE,

$$\eta_t = \gamma\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma\mu - \frac{\gamma^2\sigma^2}{2}}\lambda\left(\frac{e^{\gamma\underline{v}}}{\lambda-\gamma} - e^{\frac{1}{2}\nu^2}\frac{e^{(\gamma-b)\underline{v}}}{\lambda+b-\gamma}\right)\frac{\omega}{1-\omega} + h\frac{\mathbb{E}_t^a[p]}{1-\mathbb{E}_t^a[p]}. \quad (12)$$

One way to conduct the test is to ignore learning (setting $h = 0$ in (12)) and only test whether the representative agent's belief is anchored to the true parameter value. In this case, the RE model restriction we are testing is $\omega = p$. Alternatively, we can conduct the LR test with learning. The RE model restriction in this case is $\omega = p$ and $h = 0$. Due to the small sample size, we compute the small-sample critical values for the LR tests using simulation.

The results are reported in Table 1. For Model 1 where the agent has a lot of knowledge about the true value of p_0, the rejection rate is low regardless of whether

Table 1: Rejection Rates of Likelihood Ratio Tests on Rational Expectation Hypothesis.

	Without Learning			With Learning		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
$p_0 = 1\%$	0.046	0.015	0.322	0.048	0.473	0.101
$p_0 = 20\%$	0.049	0.021	0.098	0.049	0.073	0.082

we take learning into account or not. This is consistent with the observation that the RE assumption is a good approximation for a model with learning when agents have learned a great deal about the unknown parameter from the long series of prior data. For Model 2, the rejection rate is still low when $p_0 = 20\%$. This is because the risk premium η_t is insensitive to changes in the agent's posterior belief about p. When $p_0 = 1\%$, the likelihood ratio test becomes significantly more powerful when we take learning into account. However, even in that case, we get a rejection rate of less than 50%. Similarly, for Model 3 where the agent has irrational beliefs, it is difficult to reject the RE model based on the likelihood ratio test.

The example above illustrates the problems with RE models that we would like to emphasize. Ideally, the RE model should be a good approximation to Model 1 but not to Models 2 and 3. However, with limited data, it is difficult to distinguish the three settings, and thus, even when the RE assumption is a poor approximation to the data generating process, it may still be difficult to reject it using standard statistical tests. The informativeness of the cross-equation restrictions can be used to guard against two types of problems. For an econometrician, the danger of imposing RE when the cross-equation restrictions appear highly informative (which in our example corresponds to the cases when p_0 is small) is that we can draw wrong inferences about the true parameter value. For a model builder, the danger is that the precise beliefs imposed on the agents inside the RE model may not be anchored to the true probability distributions, but rather picked essentially arbitrarily to fit the data. The lack of understanding of how such beliefs come about not only means the model is

incomplete as an explanation of the data, but also makes it problematic to apply the model elsewhere, for instance for welfare analysis, which could also be sensitive to parameter values.

3 Information Measures

Our measures of the informativeness of the cross-equation restrictions aim to quantify the degree of precision the econometrician can gain in estimating the parameters with and without imposing the cross-equation restrictions. We define the measures formally in this section.

Let $\mathbf{x^n} = (\mathbf{x_1}, \cdots, \mathbf{x_n})$ be the sample of variables observed by agents and the econometrician. We denote the statistical model of $\mathbf{x^n}$ as $f_{\mathbb{P}}(\mathbf{x^n}; \theta, \phi)$, or simply $\mathbb{P}_{\theta,\phi}$. The subject of interest is a subset of parameters $\theta \in \mathbb{R}^d$ from the statistical model that governs the dynamics of $\mathbf{x}_t$. In particular, θ might include parameters that resemble "dark matter", i.e., parameters that are difficult to estimate directly but have large effects on model performance. The other parameters in the model (the nuisance parameters) are in the vector ϕ.

In a rational expectations model, we assume that the agents know the true parameter values and that the equilibrium decision rules and prices are based on such knowledge. Thus, the economic model can generate additional restrictions on the dynamics of $\mathbf{x}_t$, which will aid the estimation of θ. Moreover, after imposing the economic restrictions, there could be additional data $\mathbf{y^n} = (\mathbf{y_1}, \cdots, \mathbf{y_n})$ that become informative about θ.[7] We denote the economic model for the joint distribution of $(\mathbf{x^n}, \mathbf{y^n})$ as $f_{\mathbb{Q}}(\mathbf{x^n}, \mathbf{y^n}; \theta, \phi)$, or simply $\mathbb{Q}_{\theta,\phi}$, where all the nuisance parameters (including those from the statistical model) are again collected in ϕ.

We denote the true value for θ by θ_0 and denote the true value for the nuisance

[7] The distinction between $\mathbf{x^n}$ and $\mathbf{y^n}$ is that all the available data that are informative about θ in the absence of the cross-equation restrictions are included in $\mathbf{x^n}$, and $\mathbf{y^n}$ is only informative about θ with the cross-equation restrictions. The set of $\mathbf{y^n}$ can be empty.

parameter ϕ by ϕ_0. Unless stated otherwise, we assume the true values of the nuisance parameters are known, i.e., $\phi = \phi_0$. For notational simplicity, we denote $\mathbb{Q}_\theta \equiv \mathbb{Q}_{\theta,\phi_0}$, $\mathbb{P}_\theta \equiv \mathbb{P}_{\theta,\phi_0}$, $\mathbb{Q}_0 \equiv \mathbb{Q}_{\theta_0,\phi_0}$ and $\mathbb{P}_0 \equiv \mathbb{P}_{\theta_0,\phi_0}$.

3.1 Asymptotic measure

The first measure of the information provided by the cross-equation restriction is based on the comparison between the Fisher information for the parameters of interest θ in the model with and without the cross-equation restrictions. Given a likelihood function $L(\theta; \phi_0 \,|\, \text{Data})$, the Fisher information for θ is defined as

$$\mathbf{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln L(\theta; \phi_0 \,|\, \text{Data})}{\partial\theta\partial\theta^T}\right]. \tag{13}$$

We denote the information matrix under the unrestricted model by $\mathbf{I}_{\mathbb{P}}(\theta)$. Under the model with cross-equation restrictions, the information matrix is $\mathbf{I}_{\mathbb{Q}}(\theta)$.

Fisher information is naturally linked to the information one can obtain on parameters from the data. The asymptotic variance of the maximum likelihood estimator (MLE) is given by the inverse of the Fisher information. The asymptotic variance for the posterior is also given by the inverse of the Fisher information. In Section 3.3, we discuss further the information-theoretic interpretations of the measure based on Fisher information.

To compute the Fisher information requires the knowledge of the true value of the parameters. We first define the asymptotic information ratio conditional on θ. Then, we discuss how to compute the information ratio when the econometrician does not know the true value of θ.

Definition 1. *The asymptotic information ratio is defined as:*

$$\varrho_a(\theta) = \max_{\mathbf{v}\in\mathbb{R}^d, ||\mathbf{v}||=1} \frac{\mathbf{v}'\,\mathbf{I}_{\mathbb{Q}}(\theta)\,\mathbf{v}}{\mathbf{v}'\,\mathbf{I}_{\mathbb{P}}(\theta)\,\mathbf{v}}. \tag{14}$$

The idea behind this information measure is as follows. Through the vector $\mathbf{v}$,

we search over all possible directions in the parameter space to find the maximum discrepancy between the two information matrices. In the context of maximum likelihood estimation, the Fisher information is linked to the inverse of the asymptotic variance of the estimator. Therefore, intuitively, the asymptotic information ratio is a way to compare the asymptotic variances of the maximum likelihood estimators in the model with and without cross-equation restrictions. The information ratio also has a natural sample-size interpretation. Specifically, we ask what is the minimum sample size required for the MLE of parameters of the unrestricted model to match or exceed the precision of the MLE for the model with additional cross-equation restrictions in all possible directions. Because the information matrix is proportional to the sample size n, the unrestricted model requires a sample size ϱ_a times longer than the one used by the model with cross-equation restrictions.

The asymptotic information ratio is easy to compute. It is given by the largest eigenvalue of the matrix $\mathbf{R}^{-1}\mathbf{I}_{\mathbb{Q}}(\theta)\mathbf{R}^{-1}$, where $\mathbf{R} \equiv \mathbf{I}_{\mathbb{P}}(\theta)^{\frac{1}{2}}$. Let the corresponding eigenvector be $\mathbf{e}_{max}$. Then the direction along which the asymptotic information ratio is obtained is

$$\mathbf{v}_{max} = \mathbf{R}^{-1}\mathbf{e}_{max}/||\mathbf{R}^{-1}\mathbf{e}_{max}||. \tag{15}$$

The asymptotic information ratio defined in (14) is based on Fisher information, which requires full knowledge of the likelihood function to compute. In the cases where the likelihood function is unknown or too complex to compute, the Generalized Method of Moments provides a general alternative, where a generalized information matrix for GMM (see Hansen (1982), Chamberlain (1987), Hahn, Newey, and Smith (2011)) can be used in place of the Fisher information.[8] Another benefit of using the GMM is that it provides the flexibility of focusing on specific aspects of a model that the modeler considers relevant (as indicated by the selected moment conditions) when measuring the informativeness of the cross-equation restrictions.

[8]Under the GMM framework, we consider $\mathbf{I}_M^{GMM}(\theta) := G_M^T(\theta)\Omega_M^{-1}(\theta)G_M(\theta)$, where $M = \mathbb{P}$ or $\mathbb{Q}$, $G_M(\theta) := \mathbb{E}_{M_\theta}\left[\frac{\partial}{\partial\theta}g_M(x|\theta)\right]$, $\Omega_M(\theta) := \mathbb{E}_{M_\theta}\left[g_M(x|\theta)g_M^T(x|\theta)\right]$, and $g_M(x|\theta)$ is a vector-valued function. We can estimate $\mathbf{I}_M^{GMM}(\theta)$ using corresponding sample moments.

Sometimes a model can have a large number of parameters, and the econometrician might be particularly interested in examining the information content for a subset of θ. This is easy to do because the asymptotic information ratio is based on the Fisher information matrix as opposed to its inverse. The following result shows that the asymptotic information ratio based on a subset of parameters is always smaller than based on the full set.

Proposition 1. *For any non-empty subset of the model parameters denoted by $\theta_s \subseteq \theta$, let $\varrho_a(\theta_s)$ be the asymptotic information ratio based on θ_s defined in (14). Then,*

$$\varrho_a(\theta_s) \leq \varrho_a(\theta). \tag{16}$$

Proof. Without loss of generality, we assume θ_s consists of the first d_s elements of θ. Let the i-th basis vector be $e_i \equiv (0, \cdots, \underbrace{1}_{i\text{-}th}, \cdots, 0)$. Then, from (14) we have

$$\begin{aligned}\varrho_a(\theta_s) &:= \max_{\mathbf{v}\in\mathbb{R}^{d_s}:||\mathbf{v}||=1} \frac{\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_s)\mathbf{v}}{\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta_s)\mathbf{v}} = \max_{\mathbf{v}\in\mathbb{R}^{d}:||\mathbf{v}||=1,\mathbf{v}\perp e_{d_s+1},\cdots,e_d} \frac{\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta)\mathbf{v}}{\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta)\mathbf{v}} \\ &\leq \max_{\mathbf{v}\in\mathbb{R}^{d}:||\mathbf{v}||=1} \frac{\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta)\mathbf{v}}{\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta)\mathbf{v}} = \varrho_a(\theta).\end{aligned}$$

□

The asymptotic information ratio can be used in various ways in practice. We can compute ϱ_a for a model based on a specific set of calibrated parameter values. This allows us to gauge informativeness of the cross-equation restrictions under a particular calibration. Alternatively, we may want to study the fragility of a general class of models. In that case, we recommend a two-step procedure. First, we can generate a constrained posterior distribution (by imposing the cross-equation restrictions) of the parameters θ. Second, based on the distribution for θ, we can compute the distribution of $\varrho_a(\theta)$, which shows the informativeness of the cross-equation restrictions for this general class of models.[9]

[9] In the latter case, the distribution of asymptotic information ratios will depend on finite sample

3.2 Finite-sample measure

The asymptotic measure defined in the previous section does not fully exploit the information provided by the cross-equation restrictions in a given sample. In finite samples, imposing the cross-equation restrictions not only changes the variance of an estimator of θ, but also many other aspects of its distribution. Such considerations suggest that it might be desirable to compare the entire distribution of the estimators, especially when the quality of the large-sample approximation is poor.

The Bayesian method is well suited for this purpose. We first specify an "uninformative prior" about θ, $\pi(\theta)$. A truly uninformative prior is hard to define especially in the presence of constraints. We consider the Jeffreys priors of the unconstrained model $\mathbb{P}_{\theta,\phi}$. Besides the invariance property of the Jeffreys prior, it is uninformative in the sense that under general regularity conditions it maximizes the mutual information between the data and the parameters asymptotically (e.g., Polson, 1988; Clarke and Barron, 1994), hence it serves as a reference prior. The reference prior by definition makes the statistical inference maximally dependent on the data and the model, while at the same time makes the prior least informative about the parameters in a certain information-theoretic sense. Intuitively, it represents an "objective prior" in the sense of maximizing the information discrepancy between the prior and posterior distributions of the parameters. See Bernardo (1979, 2005) and Berger, Bernardo, and Sun (2009) for more discussion on the reference prior.

We use the uninformative prior to form the posterior distribution of θ in the unconstrained model. We assume the same prior for the model with cross-equation restrictions. The posterior density of θ in the unrestricted model is $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$, which denotes the dependence on the posterior on the data $\mathbf{x^n}$ and the model $\mathbb{P}_\theta$. The posterior density in the constrained model is $\pi_{\mathbb{Q}}(\theta|\mathbf{x^n}, \mathbf{y^n})$, which denotes its dependence on the data $(\mathbf{x^n}, \mathbf{y^n})$ and the cross-equation restrictions generated by the model $\mathbb{Q}_\theta$.

The relative entropy (also known as the Kullback-Leibler divergence) is a standard

information because the posterior of θ depends on the finite sample.

measure of the statistical discrepancy between two probability distributions. The relative entropy between $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$ and $\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})$ is

$$\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})||\pi_{\mathbb{P}}(\theta|\mathbf{x^n})) = \int \ln\left(\frac{\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})}{\pi_{\mathbb{P}}(\theta|\mathbf{x^n})}\right)\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})d\theta\,. \tag{17}$$

Intuitively, we can think of the log posterior ratio $\ln(\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})/\pi_{\mathbb{P}}(\theta|\mathbf{x^n}))$ as a measure of the discrepancy between the two posteriors at a given θ. Then the relative entropy is the average discrepancy between the two posteriors over all possible θ, where the average is computed under the constrained posterior. According to the definition of relative entropy (see e.g. Cover and Thomas, 1991, page 231), the relative entropy $\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})||\pi_{\mathbb{P}}(\theta|\mathbf{x^n}))$ is finite if and only if the support of the posterior $\pi_{\mathbb{Q}}(\theta|\mathbf{x^n},\mathbf{y^n})$ is a subset of the support of the posterior $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$, that is, Assumption PQ in Appendix A.1 holds. Otherwise, the relative entropy is infinite.

The relative entropy is difficult to interpret directly. We will define an entropy-based information ratio that has a similar "sample size" interpretation as the asymptotic information ratio. The idea is that instead of imposing the cross-equation restrictions, we could have gained more information about θ from extra data. Our finite-sample information ratio shows how much extra data is needed if we want to gain the same amount of information from the extra data, according to the relative entropy measure, as we do from imposing the cross-equation restrictions.

Let the additional data $\tilde{\mathbf{x}}^{\mathbf{m}}$ be of sample size m, which we assume is randomly drawn from the posterior predictive distribution

$$\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}|\mathbf{x^n}) := \int \pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}|\theta)\pi_{\mathbb{P}}(\theta|\mathbf{x^n})\mathrm{d}\theta, \tag{18}$$

where $\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}|\theta)$ is the likelihood function of the unconstrained model and $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$ is the unconstrained posterior density given historical data $\mathbf{x^n}$. Then, the gain in

information from extra data is

$$\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\theta|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})) = \int \ln\left(\frac{\pi_{\mathbb{P}}(\theta|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})}{\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})}\right)\pi_{\mathbb{P}}(\theta|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})\mathrm{d}\theta\,. \tag{19}$$

The relative entropy $\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\theta|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}}))$ depends on the realization of the additional sample of data $\tilde{\mathbf{x}}^{\mathbf{m}}$. We want to find m^* such that the amount of information provided by the extra data is, *on average*, equal to the amount of information provided by the cross-equation restrictions. The average relative entropy (information gain) over possible future sample paths $\{\tilde{\mathbf{x}}^{\mathbf{m}}\}$ according to the posterior predictive distribution $\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}|\mathbf{x}^{\mathbf{n}})$ is in fact the mutual information between $\tilde{\mathbf{x}}^{\mathbf{m}}$ and θ given $\mathbf{x}^{\mathbf{n}}$:

$$\begin{aligned}\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x}^{\mathbf{n}}) &\equiv \mathbb{E}^{(\tilde{\mathbf{x}}^{\mathbf{m}}|\mathbf{x}^{\mathbf{n}})}\left[\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\theta'|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta'|\mathbf{x}^{\mathbf{n}}))\right]\\ &= \int\int \mathbf{D}_{KL}(\pi_{\mathbb{P}}(\theta'|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta'|\mathbf{x}^{\mathbf{n}}))\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}|\theta)\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})\mathrm{d}\tilde{\mathbf{x}}^{\mathbf{m}}\mathrm{d}\theta\\ &= \int \mathbb{E}^{\tilde{\mathbf{x}}^{\mathbf{m}}}_{\mathbb{P}_\theta}\left[\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\theta'|\tilde{\mathbf{x}}^{\mathbf{m}},\mathbf{x}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta'|\mathbf{x}^{\mathbf{n}}))\right]\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})\mathrm{d}\theta.\end{aligned} \tag{20}$$

Like the relative entropy, the mutual information is always positive. It is easy to check that $\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x}^{\mathbf{n}}) = 0$ when $m = 0$. Under the assumption that the prior distribution is nonsingular and the parameters in the likelihood function are well identified, and additional general regularity conditions, $\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x}^{\mathbf{n}})$ is monotonically increasing in m and converges to infinity as m increases. These properties ensure that we can find an extra sample size m that equates (approximately, due to the fact that m is an integer) $\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}}))$ with $\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x}^{\mathbf{n}})$. Thus, we define the finite-sample information measure as follows.

Definition 2. *The finite-sample information ratio is defined as*

$$\varrho_{KL}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}}) = \frac{n+m^*}{n}, \tag{21}$$

where m^* *satisfies*

$$\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}^*};\theta|\mathbf{x}^{\mathbf{n}}) \leq \mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})) \leq \mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}^*+\mathbf{1}};\theta|\mathbf{x}^{\mathbf{n}}), \tag{22}$$

with $\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}})||\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}}))$, $\mathbf{I}(\tilde{\mathbf{x}}^{m};\theta|\mathbf{x}^{\mathbf{n}})$ *defined by* (17) *and* (20)*, respectively.*

Unlike the asymptotic measure which focuses on the direction in the parameter space with the largest information discrepancy between the constrained and unconstrained model, the finite sample measure compares the overall discrepancy (based on the relative entropy) between the posterior distributions in the constrained and unconstrained model. One can also construct a finite-sample measure that emphasizes a particular direction in the parameter space using feature functions. Definition 3 defines the finite sample information ratio with respect to a feature function. Later, Theorem 3 establishes the asymptotic equivalence between the asymptotic measure and the finite sample measure with respect to an appropriate feature function.

Definition 3. *Suppose* **Assumption PQ** *and* **FF** *in Appendix A.1 hold. The finite-sample information ratio for* θ *with respect to a feature function* f *is defined as*

$$\varrho^{f}_{KL}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}}) = \frac{n+m^*_f}{n}, \tag{23}$$

where m^*_f *satisfies*

$$\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}^*_{\mathbf{f}}};f(\theta)|\mathbf{x}^{\mathbf{n}}) \leq \mathbf{D}_{KL}(\pi_{\mathbb{Q}}(f(\theta)|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}})||\pi_{\mathbb{P}}(f(\theta)|\mathbf{x}^{\mathbf{n}})) \leq \mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}^*_{\mathbf{f}}+\mathbf{1}};f(\theta)|\mathbf{x}^{\mathbf{n}}), \tag{24}$$

with $\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(f(\theta)|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}})||\pi_{\mathbb{P}}(f(\theta)|\mathbf{x}^{\mathbf{n}}))$ *being the relative entropy between the constrained and unconstrained posteriors of* $f(\theta)$ *and* $\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};f(\theta)|\mathbf{x}^{\mathbf{n}})$ *being the conditional mutual information between the additional sample of data* $\mathbf{x}^{\mathbf{m}}$ *and the transformed parameter* $f(\theta)$ *given the existing sample of data* $\mathbf{x}^{\mathbf{n}}$*.*

What feature function should we use in practice? It is possible to search within a particular parametric family (e.g., linear functions) for the feature function that

generates the largest information discrepancy.[10] However, doing so may cause $f(\theta)$ to emphasize excessively aspects of the model that are of little economic significance. A model builder likely has a better idea about which aspects of the model implications are the most relevant, which parameters or parameter configurations are key to the model's performance, and which parameters are difficult to measure from the data. Such knowledge guides one to find the feature function that can more effectively locate the potential "dark matter" inside a model. One example is a feature function that maps the entire set of parameters θ into the subset of those "dart matter" parameters. We provide such an example in the disaster risk model studied in Section 4.

Computing the finite-sample information ratio can be numerically challenging. Next, we present two useful approximation results, one for the relative entropy between the two posteriors $\pi_{\mathbb{P}}(f(\theta)|\mathbf{x^n})$ and $\pi_{\mathbb{Q}}(f(\theta)|\mathbf{x^n},\mathbf{y^n})$, one for the mutual information $\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x^n})$. The basic idea is that in some cases we can approximate the posterior with a normal density, which gives an analytical expression for the relative entropy.

Theorem 1 (Relative Entropy)**.** *Define* $\mathbf{v} \equiv \nabla f(\theta_0)$*. Let the MLE from the unconstrained model* $\mathbb{P}_\theta$ *be* $\widehat{\theta}^{\mathbb{P}}$*, and the MLE from the constrained model* $\mathbb{Q}_\theta$ *be* $\widehat{\theta}^{\mathbb{Q}}$*. Under the regularity conditions stated in Appendix A.1, we have*

$$\begin{aligned}\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(f(\theta)|\mathbf{x^n},\mathbf{y^n})||\pi_{\mathbb{P}}(f(\theta)|\mathbf{x^n})) - \frac{n}{2\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_0)^{-1}\mathbf{v}}(f(\widehat{\theta}^{\mathbb{P}}) - f(\widehat{\theta}^{\mathbb{Q}}))^2 \\ \to \frac{1}{2}\ln\frac{\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}}{\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_0)^{-1}\mathbf{v}} + \frac{1}{2}\frac{\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_0)^{-1}\mathbf{v}}{\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}} - 1/2 \quad \textit{in } \mathbb{Q}_0. \end{aligned} \tag{25}$$

Proof. A heuristic proof is in Appendix A. A formal proof is in the Internet Appendix. □

[10] Another objective method to identify the "worst-case" feature function is to consider the maximum mean discrepancy (MMD) between the constrained and unconstrained posteriors of θ. The MMD and its approximations are studied by Gretton, Borgwardt, Rasch, Schölkopf, and Smola (2012) based on the theory of reproducing kernel Hilbert spaces (RKHS). An efficient approximation method for f^* as well as an algorithm can be found in Gretton, Borgwardt, Rasch, Schölkopf, and Smola (2012).

The following approximation has been studied extensively for mutual information:[11]

$$\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x}^{\mathbf{n}}) = \frac{d}{2}\ln\frac{m}{2\pi e} + \frac{1}{2}\int \pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})\ln|\mathbf{I}_{\mathbb{P}}(\theta)|\mathrm{d}\theta + \int \pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})\ln\frac{1}{\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})}\mathrm{d}\theta + o_p(1). \tag{26}$$

To apply the approximation (25) for the relative entropy, we need the sample size n to be large. To apply the approximation (26) for the mutual information, we need the sample size m of additional data to be large. However, the approximation (26) is valid only when the observed sample size n is fixed. The following asymptotic approximation is valid for the case where n goes to infinity.

Theorem 2 (Mutual Information)**.** *Under the assumptions in Subsection A.1, if m approaches to infinity as n goes to infinity and $m/n \to \varsigma \in (0,\infty)$, it holds that*

$$\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}}; f(\theta)|\mathbf{x}^{\mathbf{n}}) - \frac{1}{2}\ln\left(\frac{m+n}{n}\right) \to 0 \quad \textit{in } \mathbb{Q}_0.$$

Proof. A heuristic proof is in Appendix A. A formal proof is in the Internet Appendix. □

Finally, the following theorem shows that the asymptotic information ratio ϱ_a and the finite sample information ratio $\varrho^f_{KL}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}})$ are asymptotically equivalent under a certain feature function.

Theorem 3 (Asymptotic Equivalence)**.** *Consider the feature function $\hat{f}$ such that $\nabla\hat{f}(\theta_0) = \mathbf{v}_{max}$, where $\mathbf{v}_{max}$, given by (15), is the direction along which the asymptotic information ratio ϱ_a is obtained. Under the regularity conditions stated in Appendix A.1, it must hold that*

$$\ln\varrho^{\hat{f}}_{KL}(\theta|\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}}) \xrightarrow{D} \ln\varrho_a + (1-\varrho_a^{-1})(\chi^2_1 - 1),$$

[11] For more details see Clarke and Barron (1990, 1994) and references therein. See also the early development by Ibragimov and Hasminskii (1973) and Efroĭmovich (1980), and the case of non-identically distributed observations by Polson (1992), among others.

where χ_1^2 *is a chi-square random variable with degrees of freedom* 1.

Proof. A heuristic proof is in Appendix A. A formal proof is in the Internet Appendix. □

3.3 Information-theoretic interpretation for ϱ_a and ϱ_{KL}

In this section, we discuss the theoretical foundation of our measures for the informativeness of cross-equation restrictions.

The relative entropy is a widely used measure of the difference between two distributions. When $\mathbb{P}_2$ is a conditional distribution of $\mathbb{P}_1$ based on extra information, e.g., some constraints or data, $D_{KL}(\mathbb{P}_2||\mathbb{P}_1)$ is a measure of the information gain. Next, the conditional mutual information $\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}};\theta|\mathbf{x}^{\mathbf{n}})$ is an information-theoretic measure quantifying the average amount of extra information about θ embedded in the extra data $\tilde{\mathbf{x}}^{\mathbf{m}}$ relative to the information about θ already in $\mathbf{x}^{\mathbf{n}}$. Thus, the finite-sample information ratio ϱ_{KL} matches the information gain from the cross-equation restrictions with the average information gain from extra data. A similar idea is adopted by Lin, Pittman, and Clarke (2007) in studying effective sample and sample size to match a certain amount of information.

The Fisher information is a statistical measure to answer the question "How hard is it to estimate distributions." For example, the Cramer-Rao lower bound is characterized by the Fisher information. But what does it mean to compare the Fisher information from two models? We answer this question using the Chernoff information. The Chernoff information gives the asymptotic geometric rate at which the detection error probability (the weighted average of the mistake probabilities in model selection based on some prior probabilities of the two models) decays as the sample size increases. Hansen (2007) refers to it as the Chernoff rate. Intuitively, it measures "How hard is model detection based on the data." We will show that the asymptotic information ratio is the ratio of two Chernoff rates, one computed without imposing the cross-equation restrictions, one with. Thus, the asymptotic measure

quantifies the informativeness of the cross-equation restrictions by asking how much they increase our ability to distinguish alternative models.

Consider a model with density $p(x|\theta_0, \phi_0)$ and an alternative model with density $p(x|\theta_{\mathbf{v}}, \phi_0)$. Assume the densities are absolutely continuous to each other. The Chernoff information between the two models is defined as (see, e.g., Cover and Thomas (1991)):

$$C^*(p(x|\theta_{\mathbf{v}}, \phi_0) \ : \ p(x|\theta_0, \phi_0)) := -\ln \min_{\alpha \in [0,1]} \int_{\mathcal{X}} p(x|\theta_0, \phi_0)^{\alpha} p(x|\theta_{\mathbf{v}}, \phi_0)^{1-\alpha} \mathrm{d}x. \tag{27}$$

We discuss the Chernoff information further in Appendix B.2. Under a set of mild sufficient conditions which are trivially implied by the regularity conditions stated in Appendix A.1, we derive the following relationship between the asymptotic measure and the Chernoff information.

Proposition 2. *Consider the product probability measures* $p(x_1, \cdots, x_n|\theta, \phi_0) := \Pi_{i=1}^n p(x_i|\theta, \phi_0)$ *and* $q(x_1, \cdots, x_n|\theta, \phi_0) := \Pi_{i=1}^n q(x_i|\theta, \phi_0)$ *with* $\theta \in \Theta$. *Assume* Θ *is compact and* $\theta_0 \in Int(\Theta)$, *and that the densities are absolutely continuous to each other. Suppose the density function* $p(x|\theta, \phi_0)$ *and* $q(x|\theta, \phi_0)$ *are continuously differentiable in* θ *for almost every* x *under* $p(x|\theta_0, \phi_0)$. *If the elements of the Fisher Information matrixes* $\mathbf{I}_{\mathbb{P}}(\theta)$ *and* $\mathbf{I}_{\mathbb{Q}}(\theta)$ *are well defined and continuous in* θ, *then*

$$\varrho_a(\theta) = \lim_{n \to \infty} \max_{\mathbf{v} \in \mathbb{R}^d, ||\mathbf{v}||=1} \frac{C^*(q(x_1, \cdots, x_n|\theta_{\mathbf{v}}, \phi_0) \ : \ q(x_1, \cdots, x_n|\theta, \phi_0))}{C^*(p(x_1, \cdots, x_n|\theta_{\mathbf{v}}, \phi_0) \ : \ p(x_1, \cdots, x_n|\theta, \phi_0))}, \tag{28}$$

where $\theta_{\mathbf{v}} = \theta_0 + n^{-\frac{1}{2}} h\mathbf{v}$ *with* $||\mathbf{v}|| = 1$, $h \in \mathbb{R}_+$, *and* n *is the sample size.*

Proof. See Appendix B.1. □

The cross-equation restrictions increase the efficiency of parameter estimation, which makes it is easier to distinguish a model $q(x_1, \cdots, x_n|\theta, \phi_0)$ from those in its neighborhood, $q(x_1, \cdots, x_n|\theta_{\mathbf{v}}, \phi_0)$. This is reflected in a larger Chernoff rate within the class of the constrained models. The ratio between the two Chernoff rates is the largest in the direction $\mathbf{v}_{max}$, meaning that the cross-equation restrictions are the

most helpful in distinguishing between models along the $\mathbf{v}_{max}$ direction.

Finally, because the Chernoff rate is proportional to sample size, in order to match or exceed the detection error probabilities for the constrained model in all directions, one must increase the sample size for the unconstrained model by a factor of ϱ_a.

4 Disaster Risk Model Revisited

In this section, we apply our information measures to the rare disaster model introduced in Section 2. Rare economic disasters are a good example of "dark matter" in asset pricing. Statistically, one cannot rule out the existence of rare events even if they have never occurred in a limited sample of macroeconomic data. Yet, disasters can have significant impact on asset prices if they are sufficiently large in size.[12]

We consider the model under the rational expectations assumption. The expected excess log return in a non-disaster state is given in (8), which is restated below:

$$\eta \approx \gamma\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma\mu - \frac{\gamma^2\sigma^2}{2}}\lambda\left(\frac{e^{\gamma\underline{v}}}{\lambda-\gamma} - e^{\frac{1}{2}\nu^2}\frac{e^{(\gamma-b)\underline{v}}}{\lambda+b-\gamma}\right)\frac{p}{1-p}.$$

The fact that the risk premium η explodes as λ approaches γ is a crucial feature for our analysis. On the one hand, it shows that no matter how rare the disaster is, we can generate an arbitrarily large risk premium η by making the average disaster size sufficiently large (λ sufficiently small). As we will show later, these "extra rare and large" disasters are particularly difficult to rule out based on standard statistical tests. On the other hand, as the risk premium explodes, it becomes extremely sensitive to small changes in λ (its first derivative with respect to λ also explodes). This feature implies that the value of λ has to be picked "exactly right" in order to generate a risk premium consistent with the data, which makes the models based on "extra rare and large" disasters particularly fragile according to our information criteria.

[12] See the early work by Rietz (1988), and recent developments by Barro (2006), Longstaff and Piazzesi (2004), Gabaix (2012), Gourio (2012), Martin (2012), among others.

Equation (8) provides the cross-equation restriction between the process of consumption growth g_t, the disaster state z_t, and the excess log return of the market portfolio r_t. Next, we discuss how to measure the informativeness of this restriction based on the asymptotic and finite-sample information ratios.

4.1 Information ratios

We first compute the asymptotic information ratio. Since the parameters that are the most difficult to estimate directly from the data are the probability of disasters p and the disaster size parameter λ, it is natural to focus on the informativeness of the asset pricing constraint about these two parameters. Hence, we take $\theta = (p, \lambda)$ to be the parameters of interest and treat $\phi = (\mu, \sigma, \eta, \tau, \rho, \nu)$ as the nuisance parameters. In this model, the asymptotic information ratio on θ can be computed analytically.

Proposition 3. *The asymptotic information ratio for* (p, λ) *is*

$$\varrho_a(p, \lambda) = 1 + \frac{p\Delta(\lambda)^2 + p(1-p)\lambda^2\dot{\Delta}(\lambda)^2}{(1-\rho^2)\tau^2(1-p)^2} e^{2\gamma\mu - \gamma^2\sigma^2}, \tag{29}$$

where

$$\Delta(\lambda) := \lambda\left(\frac{e^{\gamma\underline{v}}}{\lambda-\gamma} - \frac{e^{(\gamma-b)\underline{v}}}{\lambda-\gamma+b}e^{\nu^2/2}\right), \tag{30}$$

and its derivative

$$\dot{\Delta}(\lambda) = -\frac{e^{\gamma\underline{v}}\gamma}{(\lambda-\gamma)^2} + \frac{e^{(\gamma-b)\underline{v}}(\gamma-b)}{(\lambda-\gamma+b)^2}e^{\nu^2/2}. \tag{31}$$

The asymptotic information ratio is obtained along the direction

$$\mathbf{v}_{max} = \left(\sqrt{p(1-p)}, \frac{\lambda^2\sqrt{\frac{1-p}{p}}\dot{\Delta}(\lambda)}{\Delta(\lambda)}\right)^T. \tag{32}$$

Proof. See Appendix C.1. □

Through (32) we can see that the direction in which the asset pricing constraint is the most informative depends on the frequency and size of disasters. If the disasters

are large and extremely rare, that is, p is small and λ is close to γ, then $\dot{\Delta}(\lambda)$ is large relative to $\Delta(\lambda)$, and the extra information provided by the asset pricing constraint will be almost entirely on the disaster size parameter λ. If the jumps in consumption are small and relatively frequent, that is, both p and λ are large, then the asset pricing constraint can become more informative about p. The direction in which asset prices provide no extra information is

$$\mathbf{v}_{min} = \left(\sqrt{p\,(1-p)}, -\frac{\Delta(\lambda)}{\sqrt{p\,(1-p)}\dot{\Delta}\,(\lambda)} \right)^{T}. \tag{33}$$

Next, we construct the Bayesian information ratio based on the unconstrained and constrained posteriors of the parameters $\theta = (p, \lambda)$ in a finite sample. We appeal to the Jeffreys prior of the model without asset pricing constraint as the econometrician's prior. Given the likelihood function in (68), the parameters are mutually independent under the Jeffreys prior and their probability density functions (PDFs) are explicitly specified in Table 2. Note that the prior $\pi(\sigma^2)\pi(\tau^2)\pi(\rho)$ is in fact the Jeffreys prior for Σ, that is,

$$\pi(\Sigma) \propto |\Sigma|^{-(d+1)/2} \quad \text{with } d = 2.$$

The posterior distribution of θ and the nuisance parameters ϕ are given in equations (82) and (83) in Appendix C.2.2. The unconstrained posterior distribution has an explicit analytical expression because Jeffreys priors are conjugate for the unconstrained model.

The constrained likelihood function (given in (70)) is "nonstandard" when we impose equality and inequality constraints on the parameters. Given the independent reference priors specified in Table 2 and the "nonstandard" likelihood function, not only the analytical form of the posterior density function becomes inaccessible, but also the traditional Monte Carlo methods designed to draw i.i.d. samples from the posterior become inefficient. For simulations based on a "nonstandard" likelihood function, one

Table 2: Independent Jeffreys/Reference Priors for Parameters

Parameters	**Prior PDF** (up to a constant)
p	$p^{-1/2}(1-p)^{-1/2}$
λ	$\lambda^{-1}\mathbf{1}_{(\lambda>0)}$
μ	$\mathbf{1}_{(-\infty<\mu<+\infty)}$
σ	$\sigma^{-2}\mathbf{1}_{(\sigma>0)}$
η	$\mathbf{1}_{(-\infty<\eta<+\infty)}$
τ	$\tau^{-2}\mathbf{1}_{(\tau>0)}$
ρ	$(1-\rho^2)^{-1}\mathbf{1}_{(-1<\rho<1)}$
ν	$\nu^{-2}\mathbf{1}_{(\nu>0)}$

of the general methods is the Approximate Bayesian Computation (ABC).[13] One issue with applying the conventional ABC method to our disaster risk model is the lack of efficiency when the priors are flat. Given the specific structure of our problem, we propose a tilted ABC method to boost the speed of our simulation. The details of the procedure are in Appendix C.2.2.

We calculate $\varrho_{KL}(\theta)$ based on the relative entropy between unconstrained posterior $\pi_{\mathbb{P}}(\theta|\mathbf{g^n},\mathbf{z^n})$ and constrained posteriors $\pi_{\mathbb{Q}}(\theta|\mathbf{g^n},\mathbf{z^n},\mathbf{r^n})$, respectively. To map into the notation of Section 3, we have $\mathbf{x^n}=\{\mathbf{g^n},\mathbf{z^n}\}$ and $\mathbf{y^n}=\{\mathbf{r^n}\}$. In general, the analytical form of the relative entropy is not available. Utilizing a large amount of simulated data from the two distributions, we estimate the relative entropy accurately using the K-Nearest-Neighbor (KNN) method (see e.g. Wang, Kulkarni, and Verdú, 2009). In addition, the average relative entropy between $\pi_{\mathbb{P}}(\theta|\mathbf{g^n},\mathbf{z^n})$ and $\pi_{\mathbb{P}}(\theta|\mathbf{g^n},\mathbf{z^n},\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})$, that is the mutual information $\mathbf{I}(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}};\theta|\mathbf{g^n},\mathbf{z^n})$, has a nearly analytical formula, which we derive in Appendix C.2.1. Using these results, compute the finite-sample information ratio $\varrho_{KL}(\theta)$ based on (21).

[13]For general introduction to the ABC method, see Blum (2010) and Fearnhead and Prangle (2012), among others.

4.2 Quantitative analysis

We now use the information ratios to study the robustness of the class of disaster risk models introduced above. We use annual real per-capita consumption growth (nondurables and services) from the NIPA and annual excess log returns of the market portfolio from CRSP for the period of 1929 to 2011.

To illustrate the fragility of the model, we plot in Figure 3 the 95% and 99% confidence regions for (p, λ) based on the unconstrained likelihood function (68). The maximum likelihood estimates are $(\widehat{p}_{MLE}, \widehat{\lambda}_{MLE}) = (0.0122, 78.7922)$, which is represented by the dot in the middle of the confidence regions. As the graph shows, the likelihood function is very flat along the direction of λ when disaster probability p is small. This feature is particularly relevant. It means that withe p sufficiently small, we cannot reject models with extremely severe disasters (small λ) using the consumption data alone. Based on this criterion, we refer to a calibration with the pair of (p, λ) that is within the 95% confidence region as an "acceptable calibration".[14]

In Figure 3 we also plot the equity premium isoquants for different levels of relative risk aversion: lines with the combinations of p and λ required to match the average equity premium of 5.89% for a given γ. The fact that these lines all cross the 95% confidence region demonstrates that even for very low risk aversion (say $\gamma = 3$), there exist many combinations of p and λ that not only match the observed equity premium, but also are "consistent with the macro data" (they are "acceptable").

While it is difficult to distinguish among these calibrations using standard statistical tools, we can use the asymptotic information ratio to determine the informativeness of the asset pricing constraint under different model calibrations, which in turn can help us gauge the robustness of the calibrated models. In particular, we focus on four different calibrations, as denoted by the four points located at the intersections of the equity premium isoquants ($\gamma = 3, 24$) and the boundary of the 95% confidence

[14] Julliard and Ghosh (2012) estimate the consumption Euler equation using the empirical likelihood method and show that the model requires a high level of relative risk aversion to match the equity premium. Their empirical likelihood criterion rules out any large disasters that have not occurred in the sample, hence requiring the model to generate high equity premium using moderate disasters.

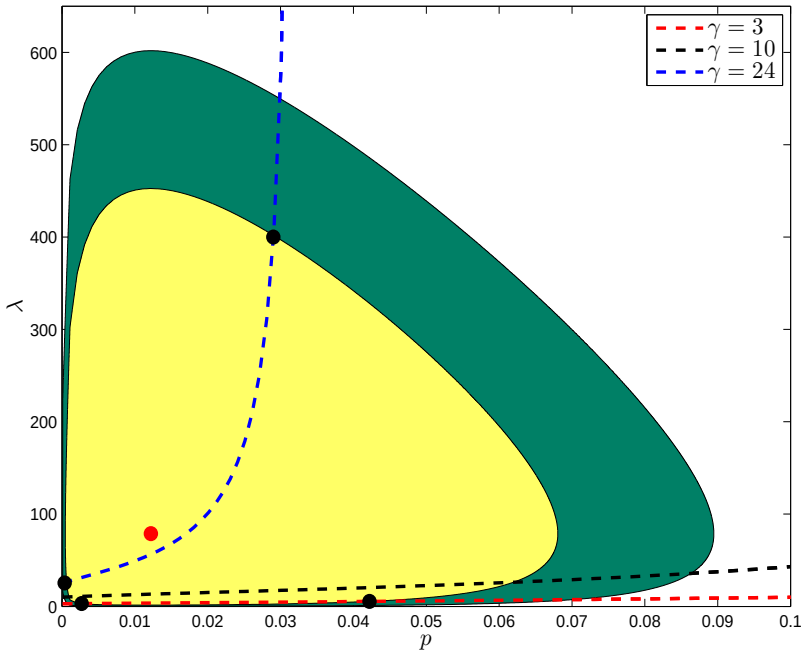


Figure 3: The 95% and 99% confidence regions of (p, λ) for the unconstrained model and the equity premium isoquants implied by the asset pricing constraint for $\gamma = 3, 10, 24$.

region (see Figure 3). For $\gamma = 3$, the two points are $(p = 4.22\%, \lambda = 5.46)$ and $(p = 0.27\%, \lambda = 3.14)$. For $\gamma = 24$, the two points are $(p = 2.9\%, \lambda = 396.7)$ and $(p = 0.037\%, \lambda = 25.49)$. As in the interest rate model, we use the MLE confidence region plots to illustrate the asymptotic information ratio in each of the four cases. The results are presented in Figure 4.

The asymptotic information ratio not only varies greatly across calibrations with different levels of relative risk aversion γ, but also across calibrations with the same γ. For example, in Panel A, the average disaster size is 25.32% and the annual disaster probability is 4.2%. In this case, $\varrho_a(p, \lambda) = 24.47$, suggesting that we need 23.47 times extra consumption data to be able to reach the same precision in the estimation of p, λ as we do with the help of the equity premium constraint. If we raise γ from 3 to 24 while changing the annual disaster probability to 2.9% and lowering the average disaster size to 7.25%, the asymptotic information ratio drops to $\varrho_a(p, \lambda) = 1.81$. The reason is that by raising the risk aversion we are able to reduce the average disaster

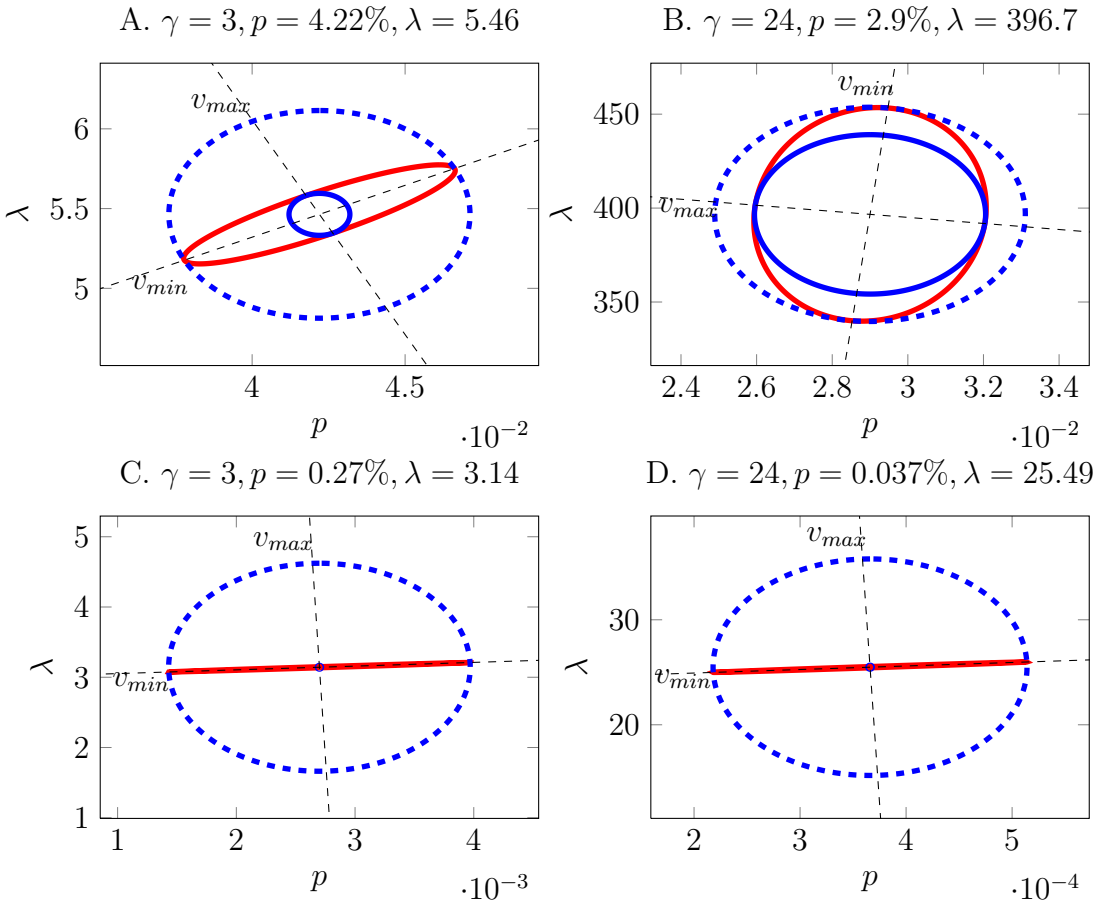


Figure 4: 95% confidence regions for the asymptotic distributions of the MLEs for four "acceptable calibrations." In Panels A through D, the asymptotic information ratios are $\rho_a(\mu, \sigma^2) = 24.47$, 1.81, 3.7×10^4, and 2.9×10^4.

size, which has dominating effect on the information ratio. Finally, Panels C and D of Figure 4 are for the calibration of "extra rare and large disasters." The impact on the asymptotic information ratio is dramatic. For $\gamma = 3$ and 24, $\varrho_a(p, \lambda)$ rises to 3.7×10^4 and 2.9×10^4, respectively.

So far we have been examining the robustness of a specific calibrated model using the asymptotic information ratio. We can also assess the robustness of a general class of models instead of a particular calibration by plotting the distribution of $\varrho_a(\theta)$ based on some "reasonable" distribution of θ. One candidate distribution is the constrained posterior distribution $\pi_{\mathbb{Q}}(\theta|\mathbf{g^n}, \mathbf{z^n}, \mathbf{r^n})$, which is discussed in Section 4.1 as part of the construction of the finite-sample information ratio. Since the constrained posterior

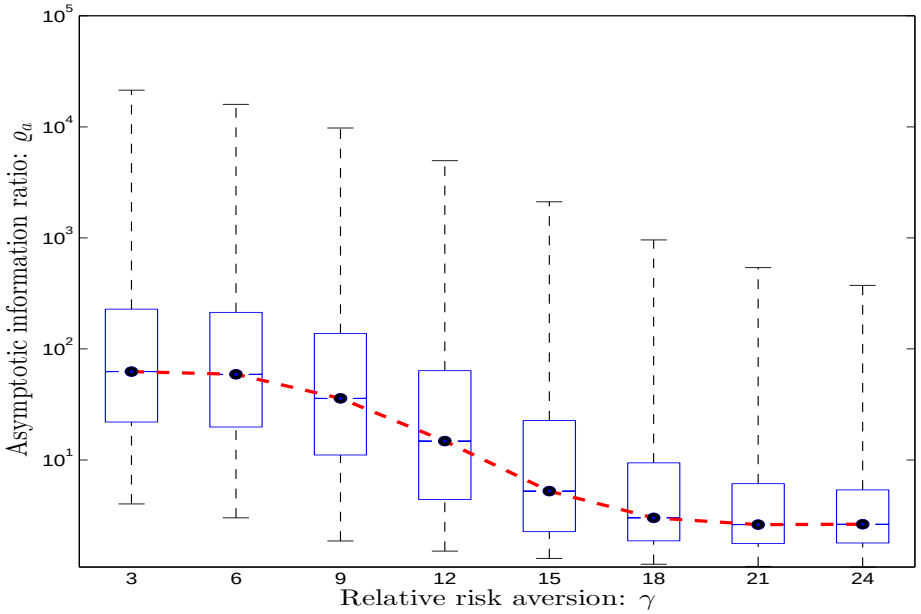


Figure 5: Distribution of asymptotic information ratios $\varrho_a(p, \lambda)$ for different levels of relative risk aversion. For each γ, the boxplot shows the $1, 25, 50, 75$ and 99-th percentile of the distribution of $\varrho_a(\theta)$ based on the constrained posterior for θ, $\pi_{\mathbb{Q}}(\theta; \gamma)$.

updates the prior $\pi(\theta)$ based on information from the data and the asset pricing constraint, it can be viewed as our "best knowledge" of the distribution of θ assuming the model constraint is valid.

We implement this idea in Figure 5. For each given γ, the boxplot shows the $1, 25, 50, 75$ and 99-th percentile of the distribution of $\varrho_a(\theta)$ based on $\pi_{\mathbb{Q}}(\theta; \gamma | \mathbf{g^n}, \mathbf{z^n}, \mathbf{r^n})$. The asymptotic information ratios are higher when the levels of risk aversion are low. For example, for $\gamma = 3$, the $25, 50$ and 75-th percentile of the distribution of $\varrho_a(p, \lambda)$ are 21.9, 62.4, and 228.0, respectively. This is because a small value of γ forces the constrained posterior for θ to place more weight on "extra rare and large" disasters, whichimposes particularly strong restrictions on the parameters (p, λ). As γ rises, the constrained posterior starts to shift its mass towards smaller disasters, which imply lower information ratios. For $\gamma = 24$, the 25-th, 50-th and 75-th percentile of the distribution of $\varrho_a(p, \lambda)$ drop to 1.8, 2.6, and 5.4, respectively.

Next, we study the finite-sample information ratio ϱ_{KL} for the disaster risk model.

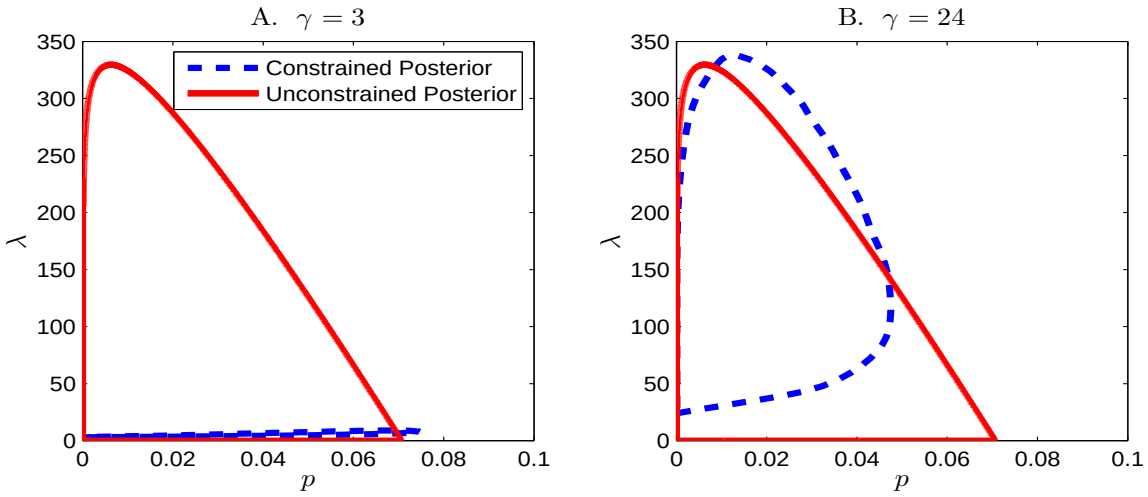


Figure 6: Unconstrained and Constrained Posterior 95% Bayesian Confidence Regions for (p, λ). In the left panel, the constrained posterior sets $\gamma = 3$. In the right panel, the constrained posterior sets $\gamma = 24$.

Since the definition of $\varrho_{KL}(\theta)$ is based on a comparison between the unconstrained posterior distribution $\pi_{\mathbb{P}}(\theta|\mathbf{g^n}, \mathbf{z^n})$ and the constrained posterior $\pi_{\mathbb{Q}}(\theta|\mathbf{g^n}, \mathbf{z^n}, \mathbf{r^n})$, in Figure 6 we illustrate their differences by plotting the 95% Bayesian confidence regions for (p, λ) according to the two posteriors. The 95% Bayesian region for the unconstrained posterior distribution is similar to the 95% confidence region for (p, λ) for the unconstrained model (see Figure 3).

The shape of the 95% Bayesian region for the constrained posterior depends on the coefficient of relative risk aversion γ. When γ is high (e.g, $\gamma = 24$), the constrained posterior is largely similar to the unconstrained posterior (see Panel B), except that it assigns lower weight to the lower right region, because these relatively frequent and large disasters are inconsistent with the equity premium constraint. For a lower level of risk aversion, $\gamma = 3$, the constrained posterior is drastically different. The only parameter configurations consistent with the equity premium constraint are those with large average disaster size, with λ close to its lower limit γ.

After computing the relative entropy via the K-Nearest-Neighbor method, we solve

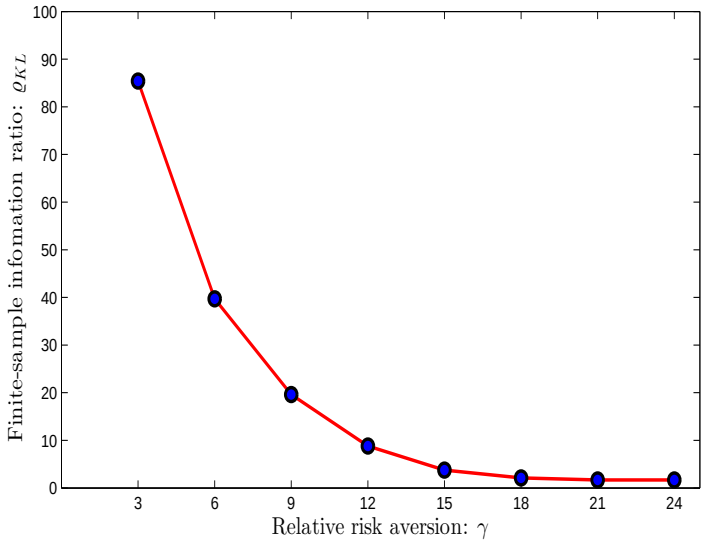


Figure 7: Finite Sample Information Ratio $\varrho_{KL}(p, \lambda)$ Conditional on γ

for m^* that satisfies

$$\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta; \gamma|\mathbf{g^n}, \mathbf{z^n}, \mathbf{r^n})||\pi_{\mathbb{P}}(\theta|\mathbf{g^n}, \mathbf{z^n})) = \mathbf{I}(\tilde{\mathbf{g}}^{\mathbf{m}^*}, \tilde{\mathbf{z}}^{\mathbf{m}^*}; \theta|\mathbf{g^n}, \mathbf{z^n}).$$

This procedure is illustrated in Figure 9 in Appendix C.2. The two horizontal lines mark the relative entropy between the unconstrained and constrained posteriors for (p, λ) for γ equal to 3 and 24. Consistent with Figure 6, the relative entropy is larger for smaller γ. The line that is rising with extra sample size m is the conditional mutual information between the extra data and the parameters, which is independent of γ. The intersections of the conditional mutual information curve with the relative entropy lines correspond to the finite-sample information ratios.

Figure 7 plots the finite-sample information ratio $\varrho_{KL}(p, \lambda)$ for a range of values of γ. Like the boxplots of the asymptotic information ratio in Figure 5, $\varrho_{KL}(p, \lambda)$ provides an overall measure of the informativeness of the asset pricing constraint for the disaster risk model. The finite-sample information ratio is again declining in γ, with $\varrho_{KL}(p, \lambda) = 85.4$ for $\gamma = 3$ and $\varrho_{KL}(p, \lambda) = 1.7$ for $\gamma = 24$. Economically, an information ratio of 85.4 implies that it would takes on average 85 times the amount

of available consumption data (close to seven thousand years) to be able to gain the same amount of extra information about model parameters as implied by the asset pricing constraint in a disaster risk model with $\gamma = 3$.

4.3 Estimating γ jointly with other model parameters

In the analysis of the rare disaster model so far, we have computed information ratios conditional on specific values of the coefficient of relative risk aversion γ. We now explicitly account for imperfect knowledge of the risk aversion parameter by the econometrician. Since the agents know their own preference parameter, it is therefore natural for asset prices to be informative about γ.

In this section we treat the risk aversion parameter γ as a part of the parameter vector to be estimated by the econometrician, which allows us to decompose the information provided by the cross-equation restrictions into information about the preference parameters and information about the remaining structural parameters.

We consider two ways to set up the joint estimation of all model parameters, including γ, in our Bayesian framework. One is to specify a prior on γ that allows for a wide range of values. Alternatively, we assume that the econometrician prefers models with low levels of risk aversion. For example, Barro (2006) states that the usual view in the finance literature is that γ is in the range of 2 to 5. Our finite-sample information ratio can accommodate such preferences. To consider the above two cases, we adopt an independent uniform prior for γ in addition to the Jeffreys priors for the other structural parameters as in Table 2. For γ, we first adopt a relatively "uninformative prior", which is uniformly distributed between 1 an 50. Next, we impose an "informative prior" on γ that is uniform between 1 and 5, which echoes the view that "a reasonable γ should not exceed 5." The ABC method can be applied to this case as well (see Appendix C.2.2 for more details).

For the case of "uninformative prior", the posterior density for γ is displayed in Panel A of Figure 8. The median value for γ is 25.81. The probability that γ is less than

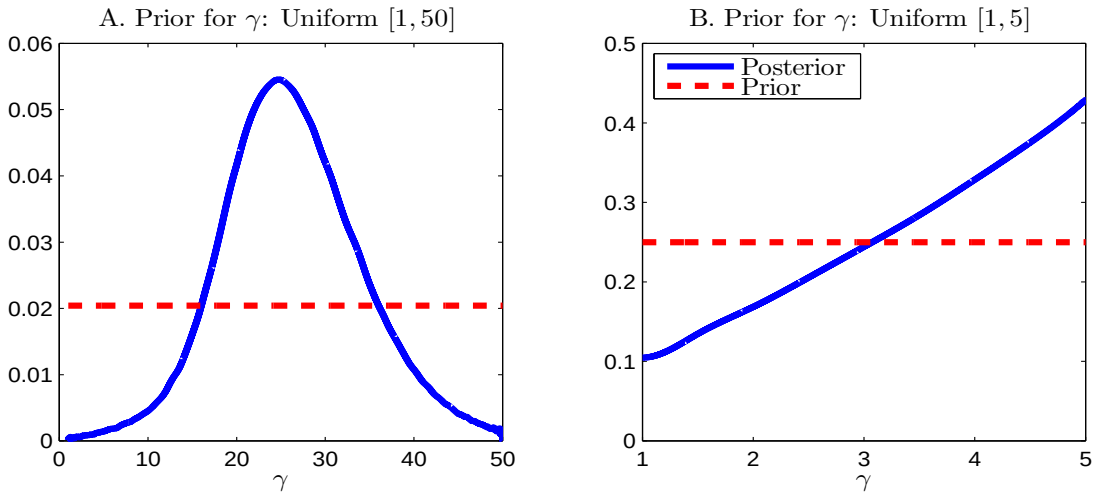


Figure 8: Posterior densities for risk-aversion parameter γ. Panel A displays the posterior and prior densities for γ when the prior is uniform on $[1, 5]$. Panel B displays the posterior and prior densities for γ when the prior is uniform on [1,50]. In this case, the posterior probability of $\gamma < 10$ is 3.93%.

10 is 3.93%. Furthermore, the finite sample information ratio $\varrho_{KL}(p, \lambda, \gamma) = 8.57$, which means that the information the econometrician gains from the asset pricing constraint is equivalent to that from 628 years of additional macro data. In this case, a major part of the information provided by asset prices is precisely on preference parameter γ, which helps the econometrician updates his prior from a uniform distribution for γ between 0 and 50 to a relatively concentrated posterior as shown in Figure 8. Another part of the information gained is on how the risk aversion parameter is correlated with the other structural parameters. Once we integrate out γ and focus exclusively on the information gained on the disaster risk parameters, we get $\varrho_{KL}(p, \lambda) = 1.27$, which means that the econometrician only needs about 22 years of extra macro data to match the information that asset prices provide on the disaster probability and average disaster size.

Next, when the econometrician holds a relatively "informative prior" that favors certain γ, the information ratios start to resemble those we computed conditional on specific values of γ. When the prior is uniform between $[1, 5]$ (see Panel B of Figure 8), the median value for γ in the posterior distribution is 3.60. The finite sample

Table 3: 95% Bayesian Confidence Intervals for Individual Parameters

Parameters	**Constrained**		**Unconstrained**
	$\mathbf{U}[1, 50]$	$\mathbf{U}[1, 5]$	
γ	[11.478, 42.772]	[1.2364, 4.9398]	
p	[0.0010, 0.0505]	[0.0053, 0.0803]	[0.0013, 0.0554]
λ	[19.320, 339.489]	[1.860, 11.614]	[1.991, 290.871]
μ	[0.0152, 0.0231]	[0.0150, 0.0227]	[0.0169, 0.0231]
σ	[0.0169, 0.0230]	[0.0169, 0.0230]	[0.0144, 0.0230]

information ratio is $\varrho_{KL}(p, \lambda, \gamma) = 184.65$, which means that the econometrician needs about 15059 years of additional macro data to match the information provided by asset prices. For $\theta = (p, \lambda)$, the finite sample information ratio is 15.9, which means that the econometrician needs about 1225 years of macro data to match the information provided by asset prices.

Finally, Table 3 reports the Bayesian posterior confidence intervals for a subset of individual parameters in the constrained and unconstrained model. Again, it is difficult to capture the notion of model fragility based on a simple comparison between the constrained and unconstrained confidence intervals for individual parameters. Only in the case with a Uniform $[0, 5]$ prior on γ did we see a sizable change in the confidence interval for the disaster size parameter λ.

5 Conclusion

Under rational expectations assumption, agents know the true probability distribution inside a model. This assumption removes the need for specifying any subjective beliefs for the agents, which simplifies and disciplines model specification and analysis. But how do we know when this assumption is acceptable and when it becomes more tenuous? In this paper, we provide new measures to systematically quantify the informational burden that a rational expectations model places on the agents.

Our methodology can be applied to rational expectations models in a wide range of areas in economics. The information measures can be used to detect model fragility based on parameter values from specific calibrations or structural estimations. They can also be used to evaluate competing classes of models that try to explain the same set of empirical phenomena. It would be interesting to formally extend the methodology to models of Bayesian learning or models with ambiguity, which we leave for future research.

References

Barro, R. J., 2006, "Rare disasters and asset markets in the twentieth century," *The Quarter Journal of Economics*, 121, 823–866.

Barro, R. J., and J. F. Ursua, 2011, "Rare Macroeconomic Disasters," NBER Working Papers 17328, National Bureau of Economic Research, Inc.

Berger, J. O., J. M. Bernardo, and D. Sun, 2009, "The formal definition of reference priors," *The Annals of Statistics*, pp. 905–938.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao, 2012, "Valid Post-Selection Inference," *Forthcoming in the Annals of Statistics*.

Bernardo, J.-M., 1979, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. Ser. B*, 41, 113–147, With discussion.

Bernardo, J. M., 2005, "Reference analysis," in *Bayesian thinking: modeling and computation*, vol. 25 of *Handbook of Statist.*, . pp. 17–90, Elsevier/North-Holland, Amsterdam.

Blum, M. G. B., 2010, "Approximate Bayesian computation: a nonparametric perspective," *J. Amer. Statist. Assoc.*, 105, 1178–1187, With supplementary material available online.

Blume, L., and D. Easley, 2010, "Heterogeneity, Selection, and Wealth Dynamics," *Annual Review of Economics*, 2, 425–450.

Campbell, J. Y., and R. J. Shiller, 1988, "Stock Prices, Earnings, and Expected Dividends," *Journal of Finance*, 43, 661–76.

Chamberlain, G., 1987, "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

Clarke, B. S., 1999, "Asymptotic normality of the posterior in relative entropy," *IEEE Trans. Inform. Theory*, 45, 165–176.

Clarke, B. S., and A. R. Barron, 1990, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, 36, 453–471.

Clarke, B. S., and A. R. Barron, 1994, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Plann. Inference*, 41, 37–60.

Collin-Dufresne, P., M. Johannes, and L. A. Lochstoer, 2013, "Parameter Learning in General Equilibrium: The Asset Pricing Implications," Working Paper.

Cover, T. M., and J. A. Thomas, 1991, *Elements of information theory* . Wiley Series in Telecommunications, John Wiley & Sons Inc., New York, A Wiley-Interscience Publication.

Efroĭmovich, S. Y., 1980, "Information contained in a sequence of observations," *Problems Inform. Transmission*, pp. 24–39.

Epstein, L. G., and M. Schneider, 2003, "Recursive multiple-priors," *Journal of Economic Theory*, 113, 1–31.

Epstein, L. G., and M. Schneider, 2010, "Ambiguity and Asset Markets," *Annual Review of Financial Economics*, 2, 315–346.

Fearnhead, P., and D. Prangle, 2012, "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation," *Journal of the Royal Statistical Society Series B*, 74, 419–474.

Gabaix, X., 2012, "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance," *The Quarterly Journal of Economics*, 127, 645–700.

Gilboa, I., and D. Schmeidler, 1989, "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

Gourio, F., 2012, "Disaster Risk and Business Cycles," *American Economic Review*, 102, 2734–66.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, 2012, "A kernel two-sample test," *J. Mach. Learn. Res.*, 13, 723–773.

Hahn, J., W. Newey, and R. Smith, 2011, "Tests for neglected heterogeneity in moment condition models," CeMMAP working papers CWP26/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Hansen, L. P., 1982, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., 2007, "Beliefs, Doubts and Learning: Valuing Macroeconomic Risk," *American Economic Review*, 97, 1–30.

Hansen, L. P., and T. J. Sargent, 1980, "Formulating and estimating dynamic linear rational expectations models," *Journal of Economic Dynamics and Control*, 2, 7–46.

Hansen, L. P., and T. J. Sargent, 1991, *Rational Expectations Econometrics*, Westview Press, Boulder, Colorado.

Hansen, L. P., and T. J. Sargent, 2001, "Robust Control and Model Uncertainty," *American Economic Review*, 91, 60–66.

Hansen, L. P., and T. J. Sargent, 2008, *Robustness*, Princeton University Press, Princeton, New Jersey.

Ibragimov, I. A., and R. Z. Hasminskii, 1973, "On the information in a sample about a parameter," *In: Proc. 2nd Internat. Symp. on Information Theory*, pp. 295 – 309.

Julliard, C., and A. Ghosh, 2012, "Can Rare Events Explain the Equity Premium Puzzle?," *Review of Financial Studies*, 25, 3037–3076.

Klibanoff, P., M. Marinacci, and S. Mukerji, 2005, "A Smooth Model of Decision Making under Ambiguity," *Econometrica*, 73, 1849–1892.

Krantz, S. G., and H. R. Parks, 2013, *The implicit function theorem* . Modern Birkhäuser Classics, Birkhäuser/Springer, New York, History, theory, and applications, Reprint of the 2003 edition.

Lehmann, E. L., and G. Casella, 1998, *Theory of point estimation* . Springer Texts in Statistics, Springer-Verlag, New York, second edn.

Lin, X., J. Pittman, and B. Clarke, 2007, "Information conversion, effective samples, and parameter size," *IEEE Trans. Inform. Theory*, 53, 4438–4456.

Longstaff, F. A., and M. Piazzesi, 2004, "Corporate earnings and the equity premium," *Journal of Financial Economics*, 74, 401–421.

Lucas, R. E., and T. J. Sargent, 1981, *Rational Expectations and Econometric Practice*, University of Minnesota Press, Minneapolis.

Martin, I., 2012, "Consumption-Based Asset Pricing with Higher Cumulants," Forthcoming, Review of Economic Studies.

Nakamura, E., J. Steinsson, R. Barro, and J. Ursa, 2012, "Crises and Recoveries in an Empirical Model of Consumption Disasters," Forthcoming, American Economic Journal: Macroeconomics.

Polson, N. G., 1988, "Bayesian Perspectives on Statistical Modelling," Ph.D. thesis, University of Nottingham.

Polson, N. G., 1992, "On the expected amount of information from a nonlinear model," *J. Roy. Statist. Soc. Ser. B*, 54, 889–895.

Rietz, T. A., 1988, "The Equity Premium: A Solution," *Journal of Monetary Economics*, 22, 117–131.

Rockafellar, R., 1970, *Convex Analysis* . Princeton mathematical series, Princeton University Press.

Saracoglu, R., and T. J. Sargent, 1978, "Seasonality and portfolio balance under rational expectations," *Journal of Monetary Economics*, 4, 435–458.

van der Vaart, A. W., 1998, *Asymptotic statistics* . , vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.

Wachter, J., 2008, "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?," NBER Working Papers 14386, National Bureau of Economic Research, Inc.

Wald, A., 1949, "Note on the Consistency of the Maximum Likelihood Estimate," *Ann. Math. Statist.*, 20, 595–601.

Wang, Q., S. R. Kulkarni, and S. Verdú, 2009, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inform. Theory*, 55, 2392–2405.

Weitzman, M. L., 2007, "Subjective Expectations and Asset-Return Puzzles," *American Economic Review*, 97, 1102–1130.

Appendix

A Heuristic Proof of Theorems

A.1 The Regularity Conditions

Assumption P

Suppose the parameter set is $\mathcal{F} = \Theta \times \Phi \subset \mathbb{R}^d \times \mathbb{R}^{d_1}$ with Θ compact. The true parameter θ_0 is an interior point of Θ. The prior is absolutely continuous with respect to the Lebesgue measure with Radon-Nykodim density $\pi_{\mathbb{P}}(\theta)$, is twice continuously differentiable and is positive on Θ.[15]

Assumption F

Suppose (θ_0, ϕ_0) is an interior point of parameter set $\mathcal{F}$. The densities $f_{\mathbb{P}}(x|\theta, \phi_0)$ and $f_{\mathbb{Q}}(x, y|\theta, \phi_0)$ are twice continuously differentiable in parameter set, for almost every x, y under $\mathbb{Q}_0$. The probability measure $f_{\mathbb{P}}(x|\theta, \phi_0)$ is the marginal distribution of the joint probability measure $f_{\mathbb{Q}}(x, y|\theta, \phi_0)$. We denote $\pi_{\mathbb{P}}(x|\theta) = f_{\mathbb{P}}(x|\theta, \phi_0)$ and $\pi_{\mathbb{Q}}(x|\theta) = f_{\mathbb{Q}}(x|\theta, \phi_0)$ by leaving out the nuisance parameter ϕ_0. For each pair of j and k, it holds that for some constant $\zeta > 0$ and large constant $C > 0$,

$$\mathbb{E}_{\mathbb{Q}_\theta} \sup_{\vartheta \in \Theta} \left| \frac{\partial^2}{\partial \vartheta_j \partial \vartheta_k} \ln \pi_{\mathbb{P}}(\mathbf{x}|\vartheta) \right|^{2+\zeta} < C, \tag{34}$$

and

$$\mathbb{E}_{\mathbb{Q}_\theta} \sup_{\vartheta \in \Theta} \left| \frac{\partial^2}{\partial \vartheta_j \partial \vartheta_k} \ln \pi_{\mathbb{Q}}(\mathbf{x}|\vartheta) \right|^{2+\zeta} < C, \tag{35}$$

and

$$\mathbb{E}_{\mathbb{Q}_\theta} \sup_{\vartheta \in \Theta} \left| \frac{\partial}{\partial \vartheta_j} \ln \pi_{\mathbb{P}}(\mathbf{x}|\vartheta) \right|^{2} < C, \tag{36}$$

and

$$\mathbb{E}_{\mathbb{Q}_\theta} \sup_{\vartheta \in \Theta} \left| \frac{\partial}{\partial \vartheta_j} \ln \pi_{\mathbb{Q}}(\mathbf{x}|\vartheta) \right|^{2} < C. \tag{37}$$

Remark 4. *There are two information matrices that typically coincide and have a*

[15] In our diaster risk model, the parameter set is not compact due to the adoption of uninformative prior. However, in that numerical example, we can truncate the parameter set at very large values which will not affect the main numerical results.

basic role in the analysis. For the model $\mathbb{P}_\theta$, these matrices are

$$\mathbf{I}_{\mathbb{P}}(\theta) \equiv \mathbb{E}_{\mathbb{Q}_\theta}\left[\frac{\partial}{\partial\vartheta}\ln\pi_{\mathbb{P}}(\mathbf{x}|\vartheta)\frac{\partial}{\partial\vartheta^T}\ln\pi_{\mathbb{P}}(\mathbf{x}|\vartheta)\Big|_{\vartheta=\theta}\right], \tag{38}$$

and

$$\mathbf{J}_{\mathbb{P}}(\theta) \equiv -\mathbb{E}_{\mathbb{Q}_\theta}\left[\frac{\partial^2}{\partial\vartheta\partial\vartheta^T}\ln\pi_{\mathbb{P}}(\mathbf{x}|\vartheta)\Big|_{\vartheta=\theta}\right]. \tag{39}$$

When the **Assumption F** *holds, we have $\mathbf{I}_{\mathbb{P}}(\theta) = \mathbf{J}_{\mathbb{P}}(\theta)$ (see e.g. Lehmann and Casella, 1998). Similarly, the assumption guarantees that $\mathbf{I}_{\mathbb{Q}}(\theta) = \mathbf{J}_{\mathbb{Q}}(\theta)$.*

Assumption KL

The Kullback-Leibler distances $\mathbf{D}_{KL}(\mathbb{P}_\theta||\mathbb{P}_\vartheta)$ and $\mathbf{D}_{KL}(\mathbb{Q}_\theta||\mathbb{Q}_\vartheta)$ are twice continuously differentiable on $\Theta\times\Theta$ with $\mathbf{J}_{\mathbb{P}}(\theta)$ and $\mathbf{J}_{\mathbb{Q}}(\theta)$ are positive definite.

Remark 5. *In fact, it follows from Assumption* **F** *that $\mathbf{D}_{KL}(\mathbb{Q}_\theta||\mathbb{Q}_\vartheta)$ and $\mathbf{D}_{KL}(\mathbb{P}_\theta||\mathbb{P}_\vartheta)$ are twice continuously differentiable in ϑ and*

$$\mathbf{J}_{\mathbb{P}}(\theta) = \frac{\partial^2}{\partial\vartheta\partial\vartheta^T}\mathbf{D}_{KL}(\mathbb{P}_\theta||\mathbb{P}_\vartheta)\mid_{\vartheta=\theta} \quad \textit{and} \quad \mathbf{J}_{\mathbb{Q}}(\theta) = \frac{\partial^2}{\partial\vartheta\partial\vartheta^T}\mathbf{D}_{KL}(\mathbb{Q}_\theta||\mathbb{Q}_\vartheta)\mid_{\vartheta=\theta}.$$

Assumption PO

For any open neighborhood $\mathcal{N}$ of θ_0, there is an $\xi_1 > 0$ and a $\xi_2 > 0$ such that

$$\mathbb{P}_0^n\mathcal{A}_{1,n}(\xi_1)^c = O\left(e^{-\xi_2 n}\right),$$

where

$$\mathcal{A}_{1,n}(\xi_1) := \left\{\int_{\mathcal{N}}\pi_{\mathbb{P}}(\vartheta)\pi_{\mathbb{P}}(\mathbf{x}^{\mathbf{n}}|\vartheta)\mathrm{d}\vartheta \geq e^{n\xi_1}\int_{\mathcal{N}^c}\pi_{\mathbb{P}}(\vartheta)\pi_{\mathbb{P}}(\mathbf{x}^{\mathbf{n}}|\vartheta)\mathrm{d}\vartheta\right\}.$$

and

$$\mathbb{Q}_0^n\mathcal{A}_{2,n}(\xi_1)^c = O\left(e^{-n\xi_2}\right).$$

where

$$\mathcal{A}_{2,n}(\xi_1) := \left\{\int_{\mathcal{N}}\pi_{\mathbb{P}}(\vartheta)\pi_{\mathbb{Q}}(\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}}|\vartheta)\mathrm{d}\vartheta \geq e^{n\xi_1}\int_{\mathcal{N}^c}\pi_{\mathbb{P}}(\vartheta)\pi_{\mathbb{Q}}(\mathbf{x}^{\mathbf{n}},\mathbf{y}^{\mathbf{n}}|\vartheta)\mathrm{d}\vartheta\right\}.$$

Remark 6. *This is a large deviation property for posterior probabilities. The condition holds under relatively mild and verifiable conditions (see e.g. Clarke, 1999, Proposition 2.1). It basically assumes that the posterior point estimator is consistent*

since it is equivalent to the following condition:

$$\mathbb{P}_0^n \left\{ \int_{\mathcal{N}} \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) d\theta < e^{n\xi_1} \int_{\mathcal{N}^c} \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) d\theta \right\} = O\left(e^{-n\xi_2}\right), \tag{40}$$

and

$$\mathbb{P}_0^n \left\{ \int_{\mathcal{N}} \pi_{\mathbb{Q}}(\theta|\mathbf{x^n}) d\theta < e^{n\xi_1} \int_{\mathcal{N}^c} \pi_{\mathbb{Q}}(\theta|\mathbf{x^n}) d\theta \right\} = O\left(e^{-n\xi_2}\right) \tag{41}$$

According to Theorem 2.2 in Clarke and Barron (1990), we know that Assumption **P** *and Assumption* **F** *imply*

$$\mathbb{Q}_0^n \left\{ 1 < e^{n\xi_1} \int_{\mathcal{N}^c} \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) \right\} = o\left(\frac{1}{n}\right), \tag{42}$$

and

$$\mathbb{Q}_0^n \left\{ 1 < e^{n\xi_1} \int_{\mathcal{N}^c} \pi_{\mathbb{Q}}(\theta|\mathbf{x^n}) \right\} = o\left(\frac{1}{n}\right). \tag{43}$$

Assumption PQ

The probability measure defined by the density $\pi_{\mathbb{Q}}(\theta|\mathbf{x^n}, \mathbf{y^n})$ is dominated by the probability measure defined by the density $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$, for almost every $\mathbf{x^n}, \mathbf{y^n}$ under $\mathbb{Q}_0$.

Assumption MLE

The Maximum Likelihood Estimators (MLEs) $\widehat{\theta^{\mathbb{P}}}$ from the model $\mathbb{P}_\theta$ and $\widehat{\theta^{\mathbb{Q}}}$ from the model $\mathbb{Q}_\theta$ exist and are consistent for each $\theta \in \Theta$.

Remark 7. *The standard sufficient conditions that guarantee* **Assumption MLE** *can be found, for example, Wald (1949), among many others. Combined with* **Assumption F***, we can show the asymptotic normality of the MLEs:*

$$\sqrt{n}(\widehat{\theta^{\mathbb{P}}} - \theta) \xrightarrow{D} N(0, \mathbf{I}_{\mathbb{P}}(\theta)^{-1}) \quad \textit{under } \mathbb{P}_\theta$$

and

$$\sqrt{n}(\widehat{\theta^{\mathbb{Q}}} - \theta) \xrightarrow{D} N(0, \mathbf{I}_{\mathbb{Q}}(\theta)^{-1}) \quad \textit{under } \mathbb{Q}_\theta.$$

And, the MLEs are asymptotically efficient; see, for example, van der Vaart (1998, § 5.5-5.6) and Lehmann and Casella (1998).

Assumption H

Define

$$H_{\mathbb{Q}}(\vartheta|\theta) := \int \pi_{\mathbb{Q}}(x, y|\theta) \ln \frac{1}{\pi_{\mathbb{Q}}(x, y|\vartheta)} \mathrm{d}x\mathrm{d}y, \tag{44}$$

and

$$H_{\mathbb{P}}(\vartheta|\theta) := \int \pi_{\mathbb{P}}(x|\theta) \ln \frac{1}{\pi_{\mathbb{P}}(x|\vartheta)} \mathrm{d}x. \tag{45}$$

Define the sample correspondences as

$$\widehat{H}_{\mathbb{Q},n}(\vartheta|\theta) := \frac{1}{n} \sum_{i=1}^{n} \ln \frac{1}{\pi_{\mathbb{Q}}(x_i, y_i|\vartheta)}, \quad \text{for } (x_i, y_i) \sim \mathbb{Q}_\theta, \tag{46}$$

and

$$\widehat{H}_{\mathbb{P},n}(\vartheta|\theta) := \frac{1}{n} \sum_{i=1}^{n} \ln \frac{1}{\pi_{\mathbb{P}}(x_i|\vartheta)}, \quad \text{for } x_i \sim \mathbb{P}_\theta. \tag{47}$$

Assume

$$\widehat{H}_{\mathbb{Q},n}(\theta|\theta_0) \to H_{\mathbb{Q}}(\theta|\theta_0) \text{ in } L^1(\mathbb{Q}_{\theta_0}) \quad \text{uniformly in } \theta \in \Theta,$$

and

$$\widehat{H}_{\mathbb{P},n}(\theta|\theta_0) \to H_{\mathbb{P}}(\theta|\theta_0) \text{ in } L^1(\mathbb{P}_{\theta_0}) \quad \text{uniformly in } \theta \in \Theta.$$

Remark 8. *This condition is also adopted by Lin, Pittman, and Clarke (2007) to guarantee the asymptotic approximation for the relative entropy.*

Assumption ID

The parametric family of joint distributions $\mathbb{Q}_{\theta,\phi}$ is sound, that is, the convergence of a sequence of parameter values is equivalent to the weak convergence of the distribution they index:

$$(\theta, \phi) \to (\theta_0, \phi_0) \Leftrightarrow \mathbb{Q}_{\theta,\phi} \to \mathbb{Q}_{\theta_0,\phi_0}. \tag{48}$$

Of course, it also holds that

$$\theta \to \theta_0 \Leftrightarrow \mathbb{P}_{\theta,\phi_0} \to \mathbb{P}_{\theta_0,\phi_0}. \tag{49}$$

Remark 9. *This assumption is a weak identifiability condition which implies that* $\theta_1 \neq \theta_2 \Rightarrow \mathbb{Q}_{\theta_1,\phi_0} \neq \mathbb{Q}_{\theta_2,\phi_0}$.

Assumption FF

The feature function $f: \Theta \to \mathbb{R}^{d'}$ with $1 \leq d' \leq d$ is twice continuously differentiable. We write $f \equiv (f_1, \cdots, f_{d'})$. We assume that there exist $d - d'$ twice continuously differentiable functions $f_2, \cdots, f_d$ on Θ such that $F = (f_1, f_2, \cdots, f_d) : \Theta \to \mathbb{R}^d$ is a one-to-one mapping (i.e. injection). Then, F is invertible and $F(\Theta)$ is also compact.

Remark 10. *A simple sufficient condition for Assumption* **FF** *to hold for* $d' = 1$ *is that* f *is a proper and twice continuously differentiable function on* $\mathbb{R}^d$ *and* $\frac{\partial f(\theta)}{\partial \theta_{(1)}} > 0$

at each $\theta \in \mathbb{R}^d$. In this case, we can simply choose $f_k(\theta) \equiv \theta_{(k)}$ for $k = 2, \cdots, d$. Then, the Jacobian determinant of F is nonzero at each $\theta \in$ and F is proper and twice differentiable differential mapping $\mathbb{R}^d \to \mathbb{R}^d$. According to the Hadamard's Global Inverse Function Theorem (see e.g. Krantz and Parks, 2013), we know that F is a one-to-one mapping.

A.2 Heuristic Proof

The following is the intuition for the proof. A formal proof is in the Internet Appendix. Without loss of generality, we assume that $\theta \in \mathbb{R}$. Under some general regularity conditions (e.g. in Appendix A.1), suppose $\widehat{\theta}^{\mathbb{P}}$ and $\widehat{\theta}^{\mathbb{Q}}$ are the MLEs for models $\mathbb{P}_0$ and $\mathbb{Q}_0$ respectively, then both the unconstrained posterior $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$ and the constrained posterior $\pi_{\mathbb{Q}}(\theta|\mathbf{x^n}, \mathbf{y^n})$ are asymptotically normal with means $\widehat{\theta}^{\mathbb{P}}$ and $\widehat{\theta}^{\mathbb{Q}}$ and variances $\Sigma_n^{\mathbb{P}}(\widehat{\theta}^{\mathbb{P}}) \equiv n^{-1}\mathbf{I}_{\mathbb{P}}^{-1}(\theta_0)$ and $\Sigma_n^{\mathbb{Q}}(\widehat{\theta}^{\mathbb{Q}}) \equiv n^{-1}\mathbf{I}_{\mathbb{Q}}^{-1}(\theta_0)$, respectively. Thus, when n is large, the relative entropy between the two posterior densities can be approximated by

$$
\begin{aligned}
&\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta|\mathbf{x^n}, \mathbf{y^n})||\pi_{\mathbb{P}}(\theta|\mathbf{x^n})) \\
&= \frac{1}{2}\left(\frac{\mathbf{I}_{\mathbb{P}}(\theta_0)}{\mathbf{I}_{\mathbb{Q}}(\theta_0)} + n\mathbf{I}_{\mathbb{P}}(\theta_0)\left(\widehat{\theta}^{\mathbb{P}} - \widehat{\theta}^{\mathbb{Q}}\right)^2 - \ln\left(\frac{\mathbf{I}_{\mathbb{P}}(\theta_0)}{\mathbf{I}_{\mathbb{Q}}(\theta_0)}\right) - 1\right) + o_p(1) \qquad (50) \\
&= \frac{1}{2}\ln\varrho_a + \frac{1}{2}\left(\varrho_a^{-1} - 1\right) + n\mathbf{I}_{\mathbb{P}}(\theta_0)\left(\widehat{\theta}^{\mathbb{P}} - \widehat{\theta}^{\mathbb{Q}}\right)^2 + o_p(1) \\
&\xrightarrow{D} \frac{1}{2}\ln\varrho_a + \frac{1}{2}(1 - \varrho_a^{-1})(\chi_1^2 - 1). \qquad (51)
\end{aligned}
$$

where $\varrho_a = \mathbf{I}_{\mathbb{P}}(\theta_0)/\mathbf{I}_{\mathbb{Q}}(\theta_0)$ in the univariate case. A similar approximation to (50) is also used in Lin, Pittman, and Clarke (2007) which also uses effective sample size to quantify the amount of information. This is the heuristic proof of Theorem 1.

Next, consider the extra information gained when increasing the sample size of the data from n to $n + m$. By Definition 2, we know that

$$
\begin{aligned}
\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}}; \theta|\mathbf{x^n}) &= \int_{\mathcal{X}^m \times \Theta} \pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \theta|\mathbf{x^n}) \ln \frac{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x^n}|\theta)\pi_{\mathbb{P}}(\mathbf{x^n})}{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x^n})\pi_{\mathbb{P}}(\mathbf{x^n}|\theta)} \mathrm{d}\tilde{\mathbf{x}}^{\mathbf{m}}\mathrm{d}\theta \\
&= \int_{\mathcal{X}^m \times \Theta} \pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \theta|\mathbf{x^n}) \ln \frac{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x^n}|\theta)}{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x^n})} \mathrm{d}\tilde{\mathbf{x}}^{\mathbf{m}}\mathrm{d}\theta + \int_{\Theta} \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) \ln \frac{\pi_{\mathbb{P}}(\mathbf{x^n})}{\pi_{\mathbb{P}}(\mathbf{x^n}|\theta)}. \qquad (52)
\end{aligned}
$$

We know that, when m and n go to infinity, we have for every $\theta \in \Theta$,

$$\ln \frac{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x}^{\mathbf{n}}|\theta)}{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x}^{\mathbf{n}})} + \frac{1}{2}\tilde{S}'_{m+n}\mathbf{I}_{\mathbb{P}}(\theta)^{-1}\tilde{S}_{m+n} - \frac{1}{2}\ln\frac{m+n}{2\pi} \xrightarrow{L_1(\mathbb{P}_\theta)} \ln\frac{1}{\pi_{\mathbb{P}}(\theta)} + \frac{1}{2}\ln|\mathbf{I}_{\mathbb{P}}(\theta)|, \tag{53}$$

and

$$\ln \frac{\pi_{\mathbb{P}}(\mathbf{x}^{\mathbf{n}}|\theta)}{\pi_{\mathbb{P}}(\mathbf{x}^{\mathbf{n}})} + \frac{1}{2}S'_{n}\mathbf{I}_{\mathbb{P}}(\theta)^{-1}S_{n} - \frac{1}{2}\ln\frac{n}{2\pi} \xrightarrow{L_1(\mathbb{P}_\theta)} \ln\frac{1}{\pi_{\mathbb{P}}(\theta)} + \frac{1}{2}\ln|\mathbf{I}_{\mathbb{P}}(\theta)|, \tag{54}$$

where

$$S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln\pi_{\mathbb{P}}(x_i|\theta),$$

$$S_m := \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\frac{\partial}{\partial\theta}\ln\pi_{\mathbb{P}}(\tilde{x}_i|\theta),$$

and

$$\tilde{S}_{m+n} := \sqrt{\frac{n}{m+n}}S_n + \sqrt{\frac{m}{m+n}}S_m. \tag{55}$$

Equations (53) and (54) hold under general regularity conditions (see Appendix A.1). For more detailed discussion on the approximation results above, see Clarke and Barron (1990, 1994) and references therein.

According to the Markov inequality and the approximation in (53), if Θ is compact in $\mathbb{R}$, it follows that

$$\begin{aligned}
&\int_{\mathcal{X}^m\times\Theta}\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \theta|\mathbf{x}^{\mathbf{n}})\ln\frac{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x}^{\mathbf{n}}|\theta)}{\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \mathbf{x}^{\mathbf{n}})}\mathrm{d}\tilde{\mathbf{x}}^{\mathbf{m}}\mathrm{d}\theta \\
&= -\int_{\mathcal{X}^m\times\Theta}\pi_{\mathbb{P}}(\tilde{\mathbf{x}}^{\mathbf{m}}, \theta|\mathbf{x}^{\mathbf{n}})\frac{1}{2}\tilde{S}'_{m+n}\mathbf{I}_{\mathbb{P}}(\theta)^{-1}\tilde{S}_{m+n}\mathrm{d}\tilde{\mathbf{x}}^{\mathbf{m}}\mathrm{d}\theta \\
&\qquad + \frac{1}{2}\ln\frac{m+n}{2\pi} + \int_{\Theta}\pi_{\mathbb{P}}(\theta|x^n)\ln\frac{1}{\pi_{\mathbb{P}}(\theta)}\mathrm{d}\theta + \frac{1}{2}\int_{\Theta}\pi_{\mathbb{P}}(\theta|x^n)\ln|\mathbf{I}_{\mathbb{P}}(\theta)|\mathrm{d}\theta + o_p(1) \\
&= -\frac{n}{2(m+n)}\int_{\Theta}\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})S'_{n}\mathbf{I}_{\mathbb{P}}(\theta)^{-1}S_n\mathrm{d}\theta - \frac{m}{2(m+n)} \\
&\qquad + \frac{1}{2}\ln\frac{m+n}{2\pi} + \int_{\Theta}\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})\ln\frac{1}{\pi_{\mathbb{P}}(\theta)}\mathrm{d}\theta + \frac{1}{2}\int_{\Theta}\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})\ln|\mathbf{I}_{\mathbb{P}}(\theta)|\mathrm{d}\theta + o_p(1)
\end{aligned} \tag{56}$$

According to the approximation in (54), if Θ is compact in $\mathbb{R}$, it follows that

$$\int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) \ln \frac{\pi_{\mathbb{P}}(\mathbf{x^n}|\theta)}{\pi_{\mathbb{P}}(\mathbf{x^n})} \mathrm{d}\theta = -\frac{1}{2}\int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) S_n' \mathbf{I}_{\mathbb{P}}(\theta)^{-1} S_n \mathrm{d}\theta + \frac{1}{2}\ln\frac{n}{2\pi} + \int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) \ln \frac{1}{\pi_{\mathbb{P}}(\theta)} \mathrm{d}\theta + \frac{1}{2}\int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) \ln |\mathbf{I}_{\mathbb{P}}(\theta)| \mathrm{d}\theta + o_p(1), \tag{57}$$

Thus, from (52), (56) and (57), it follows that

$$\mathbf{I}(\tilde{\mathbf{x}}^{\mathbf{m}}; \theta|\mathbf{x^n}) = \frac{1}{2}\ln\frac{m+n}{n} + \frac{m}{2(m+n)}\left[\int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) S_n' \mathbf{I}_{\mathbb{P}}(\theta)^{-1} S_n \mathrm{d}\theta - 1\right] + o_p(1), \tag{58}$$

and

$$\int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) S_n' \mathbf{I}_{\mathbb{P}}(\theta)^{-1} S_n \mathrm{d}\theta = n\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta^T} \ln \pi_{\mathbb{P}}(x_i|\widehat{\theta}^{\mathbb{P}})\right]^2 \int_{\mathbb{R}} \phi\left(\theta|\widehat{\theta}^{\mathbb{P}}, n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\right)(\theta - \widehat{\theta}^{\mathbb{P}})^2 \mathrm{d}\theta + o_p(1)$$

where $\phi\left(\theta|\widehat{\theta}^{\mathbb{P}}, n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\right)$ denotes the probability density function for $N(\widehat{\theta}^{\mathbb{P}}, n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0))$.

Under regularity conditions, we have

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta^T} \ln \pi_{\mathbb{P}}(x_i|\widehat{\theta}^{\mathbb{P}}) \xrightarrow{\mathbb{P}_0} \mathbf{I}_{\mathbb{P}}(\theta_0).$$

And, it holds that

$$\int_\Theta \phi\left(\theta|\widehat{\theta}^{\mathbb{P}}, n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\right)(\theta - \widehat{\theta}^{\mathbb{P}})^2 \mathrm{d}\theta = n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}.$$

Therefore, we have

$$\int_\Theta \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) S_n' \mathbf{I}_{\mathbb{P}}(\theta)^{-1} S_n \mathrm{d}\theta = 1 + o_p(1), \quad \text{under } \mathbb{Q}_0.$$

This is the heuristic proof of Theorem 2. Thus, by the Definition of $\varrho_{KL}(\theta|\mathbf{x^n}, \mathbf{y^n})$ and the Slutsky's Theorem, we have

$$\ln(\varrho_{KL}(\theta|\mathbf{x^n}, \mathbf{y^n})) \xrightarrow{D} \ln \varrho_a + (1 - \varrho_a^{-1})(\chi_1^2 - 1), \tag{59}$$

where χ_1^2 is a chi-square random variable with degrees of freedom 1. Then, we have proved Theorem 3 heuristically.

B Chernoff Rate, Fisher Information, and Detection Error Probability

B.1 Proof of Proposition 2

Let's first show the following lemma. Recall that we defined in Proposition 2 that

$$\theta_{\mathbf{v}} := \theta_0 + n^{-1/2} h\mathbf{v}.$$

Lemma 1. *Assume that Θ is compact and $\theta_0 \in \Theta$. Suppose the product probability measure with density function $p(x_1, \cdots, x_n|\theta, \phi_0) := \Pi_{i=1}^n p(x_i|\theta, \phi_0)$ where $p(x|\theta, \phi_0)$ is continuously differentiable in θ for almost every x under $p(x|\theta_0, \phi_0)$. If the elements of the Fisher Information matrix $\mathbf{I}(\theta)$ are well defined and continuous in θ, then*

$$C^*(p(x_1, \cdots, x_n|\theta_{\mathbf{v}}, \phi_0) \ : \ p(x_1, \cdots, x_n|\theta_0, \phi_0)) = \frac{h^2}{8}\mathbf{v}'\mathbf{I}(\theta_0)\mathbf{v} + o(1). \tag{60}$$

Proof. First, we have the following identity

$$\begin{aligned}
&\int_{\mathcal{X}} \left[\Pi_{i=1}^n p(x_i|\theta_0, \phi_0)\right]^{1-\alpha} \left[\Pi_{i=1}^n p(x_i|\theta_{\mathbf{v}}, \phi_0)\right]^{\alpha} \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&\qquad = \int_{\mathcal{X}} \Pi_{i=1}^n p(x_i|\theta_0, \phi_0) e^{\alpha \sum_{i=1}^n [\ln p(x_i|\theta_{\mathbf{v}}, \phi_0) - \ln p(x_i|\theta_0, \phi_0)]} \mathrm{d}x_1 \cdots \mathrm{d}x_n.
\end{aligned} \tag{61}$$

According to Theorem 7.2 and Lemma 7.6 in van der Vaart (1998), we know that the Local Asymptotic Normality (LAN) condition holds, i.e.

$$\sum_{i=1}^n \left[\ln p(X_i|\theta_{\mathbf{v}}, \phi_0) - \ln p(X_i|\theta_0, \phi_0)\right] \xrightarrow{D} N\left(-\frac{1}{2}h^2\mathbf{v}^T\mathbf{I}(\theta_0)\mathbf{v}, h^2\mathbf{v}^T\mathbf{I}(\theta_0)\mathbf{v}\right),$$

where $X_1, X_2, \cdots, X_n$ are i.i.d. from the distribution $p(x|\theta_0, \phi_0)$.

Let's define the function

$$\begin{aligned} M_n(\alpha) &\equiv \mathbb{E} e^{\alpha \sum_{i=1}^n [\ln p(X_i|\theta_{\mathbf{v}},\phi_0) - \ln p(X_i|\theta_0,\phi_0)]} \\ &= \int_{\mathcal{X}} \Pi_{i=1}^n p(x_i|\theta_0,\phi_0) e^{\alpha \sum_{i=1}^n [\ln p(x_i|\theta_{\mathbf{v}},\phi_0) - \ln p(x_i|\theta_0,\phi_0)]} \mathrm{d}x_1 \cdots \mathrm{d}x_n \end{aligned}$$

Then, we can see that the function $M_n(\alpha)$ is effectively the Moment Generating Function (MGF) of random variable $\sum_{i=1}^n [\ln p(X_i|\theta_{\mathbf{v}},\phi_0) - \ln p(X_i|\theta_0,\phi_0)]$. Thus, it follows that

$$M_n(\alpha) \to e^{-\frac{1}{2}\alpha(1-\alpha)h^2 \mathbf{v}^T \mathbf{I}_{\mathbb{P}}(\theta_0)\mathbf{v}}, \quad \text{as } n \text{ goes to infinity.}$$

We denote the Cumulant Generating Function (CGF) as

$$\Lambda_n(\alpha) = \log M_n(\alpha).$$

It is easy to show that the Cumulant Generating Function $\Lambda_n(\alpha)$ is convex for each n.

Based on the definition of Chernoff information in (27) and the identity in (61), we know that

$$\begin{aligned} & C^*(p(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ p(x_1,\cdots,x_n|\theta_0,\phi_0)) \\ &:= \max_{\alpha\in[0,1]} -\ln \int_{\mathcal{X}} [\Pi_{i=1}^n p(x|\theta_0,\phi_0)]^{\alpha} [\Pi_{i=1}^n p(x|\theta_{\mathbf{v}},\phi_0)]^{1-\alpha} \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \max_{\alpha\in[0,1]} -\Lambda_n(\alpha) \to \max_{\alpha\in[0,1]} \frac{1}{2}\alpha(1-\alpha)h^2 \mathbf{v}^T \mathbf{I}(\theta_0)\mathbf{v} = \frac{1}{8}h^2 \mathbf{v}^T \mathbf{I}(\theta_0)\mathbf{v}. \end{aligned} \tag{62}$$

The limit result in (62) is a key step in our proof. Remember that the function $-\Lambda_n(\alpha)$ is concave for each n and we know that the pointwise convergence for a sequence of concave functions implies their uniform convergence to a concave function(see e.g. Rockafellar, 1970). Finally, the uniform convergence and concavity of $-\Lambda_n(\alpha)$ guarantees the convergence of maxima of $-\Lambda_n(\alpha)$ to the maximum of $\frac{1}{2}\alpha(1-\alpha)h^2 \mathbf{v}^T \mathbf{I}(\theta_0)\mathbf{v}$, which is $\frac{1}{8}h^2 \mathbf{v}^T \mathbf{I}(\theta_0)\mathbf{v}$. □

According to Lemma 1, for any probability density functions $p(x_1,\cdots,x_n|\theta,\phi_0)$ and $q(x_1,\cdots,x_n|\theta,\phi_0)$ satisfying the conditions in Proposition 2, it holds that

$$C^*(p(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ p(x_1,\cdots,x_n|\theta_0,\phi_0)) = \frac{h^2}{8}\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta_0)\mathbf{v} + o(1) \tag{63}$$

and

$$C^*(q(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ q(x_1,\cdots,x_n|\theta_0,\phi_0)) = \frac{h^2}{8}\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_0)\mathbf{v} + o(1), \tag{64}$$

where the Fisher information matrixes are defined as

$$\mathbf{I}_{\mathbb{P}}(\theta) := \int_{\mathcal{X}} p(x|\theta,\phi_0)\left[\frac{\partial}{\partial\theta}\ln p(x|\theta,\phi_0)\right]^2 \mathrm{d}x \tag{65}$$

and

$$\mathbf{I}_{\mathbb{Q}}(\theta) := \int_{\mathcal{X}} q(x|\theta,\phi_0)\left[\frac{\partial}{\partial\theta}\ln q(x|\theta,\phi_0)\right]^2 \mathrm{d}x. \tag{66}$$

Therefore, we have

$$\begin{aligned}
&\lim_{n\to\infty}\max_{\mathbf{v}\in\mathbb{R}^d,||\mathbf{v}||=1}\frac{C^*(q(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ q(x_1,\cdots,x_n|\theta,\phi_0))}{C^*(p(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ p(x_1,\cdots,x_n|\theta,\phi_0))}\\
&= \max_{\mathbf{v}\in\mathbb{R}^d,||\mathbf{v}||=1}\lim_{n\to\infty}\frac{C^*(q(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ q(x_1,\cdots,x_n|\theta,\phi_0))}{C^*(p(x_1,\cdots,x_n|\theta_{\mathbf{v}},\phi_0) \ : \ p(x_1,\cdots,x_n|\theta,\phi_0))}\\
&= \max_{\mathbf{v}\in\mathbb{R}^d,||\mathbf{v}||=1}\lim_{n\to\infty}\frac{\frac{h^2}{8}\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_0)\mathbf{v}+o(1)}{\frac{h^2}{8}\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta_0)\mathbf{v}+o(1)}\\
&= \max_{\mathbf{v}\in\mathbb{R}^d,||\mathbf{v}||=1}\frac{\frac{h^2}{8}\mathbf{v}'\mathbf{I}_{\mathbb{Q}}(\theta_0)\mathbf{v}}{\frac{h^2}{8}\mathbf{v}'\mathbf{I}_{\mathbb{P}}(\theta_0)\mathbf{v}}\\
&= \varrho_a(\theta).
\end{aligned}$$

B.2 Chernoff rate and detection error probability

This subsection is mainly based on Section 12.9 in Cover and Thomas (1991). Assume $X_1,\cdots,X_n$ i.i.d. $\sim Q$. We have two hypothesis or classes: $Q = P_1$ with prior π_1 and $Q = P_2$ with prior π_2. The overall probability of error (detection error probability) is

$$P_e^n = \pi_1 E_1^{(n)} + \pi_2 E_2^{(n)},$$

where $E_1^{(n)}$ is the error probability when $Q = P_1$ and $E_2^{(n)}$ is the error probability when $Q = P_2$. Define the best achievable exponent in the detection error probability is

$$D^* = \lim_{n\to\infty}\min_{A_n\in\mathcal{X}^n} -\frac{1}{n}\log_2 P_e^{(n)}, \quad \text{where } A_n \text{ is the acceptance region.}$$

The Chernoff's Theorem shows that $D^* = C^*(P_1 : P_2)$. More precisely, Chernoff's Theorem states that the best achievable exponent in the detection error probability is D^*, where

$$D^* = \mathbf{D}_{KL}(P_{\alpha^*}||P_1) = \mathbf{D}_{KL}(P_{\alpha^*}||P_2),$$

with

$$P_\alpha = \frac{P_1^\alpha(x)P_2^{1-\alpha}(x)}{\int_\chi P_1^\alpha(x)P_2^{1-\alpha}(x)\mathrm{d}x}$$

and α^* is the value of α such that

$$\mathbf{D}_{KL}(P_{\alpha^*}||P_1) = \mathbf{D}_{KL}(P_{\alpha^*}||P_2) = C^*(P_1 : P_2).$$

According to the Chernoff's Theorem, intuitively, the best achievable exponent in the detection error probability is

$$P_e^{(n)} \doteq \pi_1 2^{-n\mathbf{D}_{KL}(P_{\alpha^*}||P_1)} + \pi_2 2^{-n\mathbf{D}_{KL}(P_{\alpha^*}||P_2)} = 2^{-nC^*(P_1:P_2)}. \tag{67}$$

Combining (60) and (67), we can see another way of interpreting the Fisher information ratio as the sample size ratio to achieve the same level of detection error probability asymptotically.

C Disaster risk model

C.1 Proof of Proposition 3

The joint probability density for (g, r, z) in the unconstrained model is

$$\begin{aligned}
&\pi_{\mathbb{P}}(g, r, z|\theta, \phi) = p^z(1-p)^{1-z} \\
&\times \left[\frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(g-\mu)^2}{\sigma^2} + \frac{(r-\eta)^2}{\tau^2} - \frac{2\rho(g-\mu)(r-\eta)}{\sigma\tau}\right]\right\}\right]^{1-z} \\
&\times \left[\mathbf{1}_{\{-g>\underline{v}\}}\lambda \exp\left\{-\lambda(-g-\underline{v})\right\} \frac{1}{\sqrt{2\pi}\nu} \exp\left\{-\frac{1}{2\nu^2}(r-bg)^2\right\}\right]^z.
\end{aligned} \tag{68}$$

The Fisher information matrix for (p, λ) under the unconstrained model $\mathbb{P}_{\theta,\phi}$ is

$$\mathbf{I}_{\mathbb{P}}(p, \lambda) = \begin{bmatrix} \frac{1}{p(1-p)} & 0 \\ 0 & \frac{p}{\lambda^2} \end{bmatrix}. \tag{69}$$

Next, to derive the probability density function $f_{\mathbb{Q}}(g,r,z|\theta,\phi)$ in the constrained model, we simply substitute the risk premium η in $f_{\mathbb{P}}(g,r,z|\theta,\phi)$ (given by (68)) with the asset pricing constraint (8):

$$\begin{aligned}\pi_{\mathbb{Q}}(g,r,z|\theta,\phi) &= p^z(1-p)^{1-z} \\ &\times \left[\frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(g-\mu)^2}{\sigma^2}+\frac{(r-\eta(\theta,\phi))^2}{\tau^2}-\frac{2\rho(g-\mu)(r-\eta(\theta,\phi))}{\sigma\tau}\right]\right\}\right]^{1-z} \\ &\times \left[\mathbf{1}_{\{-g>\underline{v}\}}\lambda\exp\left\{-\lambda(-g-\underline{v})\right\}\frac{1}{\sqrt{2\pi}\nu}\exp\left\{-\frac{1}{2\nu^2}(r-bg)^2\right\}\right]^{z}\mathbf{1}_{(\eta(\theta,\phi)>\underline{\eta}^*,\lambda>\gamma)}, \qquad (70)\end{aligned}$$

where

$$\eta(\theta,\phi) := \gamma\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma\mu-\frac{\gamma^2\sigma^2}{2}}\lambda\left(\frac{e^{\gamma\underline{v}}}{\lambda-\gamma}-e^{\frac{1}{2}\nu^2}\frac{e^{(\gamma-b)\underline{v}}}{\lambda+b-\gamma}\right)\frac{p}{1-p}. \qquad (71)$$

Using the notation introduced by (30) and (31), we can express the Fisher information for (p,λ) under the constrained model $\mathbb{Q}_{\theta,\phi}$ (with relative risk aversion γ) as

$$\mathbf{I}_{\mathbb{Q}}(p,\lambda;\gamma) = \begin{bmatrix} \frac{1}{p(1-p)}+\frac{\Delta(\lambda)^2}{(1-\rho^2)\tau^2}\frac{e^{2\gamma\mu-\gamma^2\sigma^2}}{(1-p)^3} & \frac{p}{(1-\rho^2)\tau^2}\frac{e^{2\gamma\mu-\gamma^2\sigma^2}}{(1-p)^2}\Delta(\lambda)\dot{\Delta}(\lambda) \\ \frac{p}{(1-\rho^2)\tau^2}\frac{e^{2\gamma\mu-\gamma^2\sigma^2}}{(1-p)^2}\Delta(\lambda)\dot{\Delta}(\lambda) & \frac{p}{\lambda^2}+\frac{\dot{\Delta}(\lambda)^2}{(1-\rho^2)\tau^2}e^{2\gamma\mu-\gamma^2\sigma^2}\frac{p^2}{1-p} \end{bmatrix}. \qquad (72)$$

Following the definition in (14), the asymptotic information ratio is the largest eigenvalue of the matrix $\mathbf{I}_{\mathbb{P}}^{-1/2}(\theta)\mathbf{I}_{\mathbb{Q}}(\theta)\mathbf{I}_{\mathbb{P}}^{-1/2}(\theta)$. In this case, the eigenvalues and eigenvectors are available in closed form. This gives us the formula for $\varrho_a(p,\lambda)$ in (29) and the worst-case direction $\mathbf{v}_{max}$ in (32). The minimum information ratio is 1, which is obtained in the direction (33).

C.2 Finite Sample Measure for the Disaster Model

In this section, we provide the details about our tilted ABC method and its implementation algorithm for simulating the constrained posterior distribution $\pi_{\mathbb{Q}}(\theta|\mathbf{g^n},\mathbf{r^n},\mathbf{z^n})$. Moreover, we also derive the analytical formula for the conditional mutual information $\mathbf{I}(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}};\theta|\mathbf{g^n},\mathbf{z^n})$ given the historical data $\mathbf{g^n},\mathbf{z^n}$.

C.2.1 Derivation of $\mathbf{I}(\widetilde{\mathbf{g}}^{\mathbf{m}}, \widetilde{\mathbf{z}}^{\mathbf{m}}; \theta|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})$

The average relative entropy between $\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})$ and $\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}, \tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}})$ over distribution $\pi_{\mathbb{P}}(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}}|\mathbf{x}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})$, that is $\mathbf{I}(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}}; \theta|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})$, has a nearly analytical formula. We derive it as follows.

$$\begin{aligned}\pi_{\mathbb{P}}(\theta|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}) = \pi_{\mathbb{P}}(p, \lambda|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}) &= \pi_{\mathbb{P}}(p|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})\pi_{\mathbb{P}}(\lambda|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}) \\ &= \pi_{\mathbb{P}}(p|\kappa_n)\pi_{\mathbb{P}}(\lambda|\Lambda_n, \kappa_n)\end{aligned}$$

where $\kappa_n = \sum_{t\le n}(1 - z_t)$ and $\Lambda_n = \sum_{t\le n} z_t(-g_t - \underline{v})$. In fact,

$$\pi_{\mathbb{P}}(p|\kappa_n) = \frac{p^{1/2+n-\kappa_n-1}(1-p)^{1/2+\kappa_n-1}}{B(1/2+n-\kappa_n, 1/2+\kappa_n)}, \tag{73}$$

and

$$\pi_{\mathbb{P}}(\lambda|\Lambda_n, \kappa_n) = \frac{1}{\Gamma(n-\kappa_n)}\Lambda_n^{n-\kappa_n}\lambda^{n-\kappa_n-1}e^{-\Lambda_n\lambda}. \tag{74}$$

With more data $\mathbf{g}^{\tilde{\mathbf{m}}}, \mathbf{z}^{\tilde{\mathbf{m}}}$, the posteriors become

$$\pi_{\mathbb{P}}(p, \lambda|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}, \tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}}) = \pi_{\mathbb{P}}(p|\kappa_n + \tilde{\kappa}_m)\pi_{\mathbb{P}}(\lambda|\Lambda_n + \tilde{\Lambda}_m, \kappa_n + \tilde{\kappa}_m)$$

with $\tilde{\kappa}_m = \sum_{1\le t\le m}(1 - \tilde{z}_t)$ and $\tilde{\Lambda}_m = \sum_{1\le t\le m}\tilde{z}_t(-\tilde{g}_t - \underline{v})$. More precisely, we have

$$\pi_{\mathbb{P}}(p|\kappa_n + \tilde{\kappa}_m) = \frac{p^{1/2+n+m-\kappa_n-\tilde{\kappa}_m-1}(1-p)^{1/2+\kappa_n+\tilde{\kappa}_m-1}}{B(1/2+n+m-\kappa_n-\tilde{\kappa}_m, 1/2+\kappa_n+\kappa_m)}, \tag{75}$$

and

$$\pi_{\mathbb{P}}(\lambda|\Lambda_n+\tilde{\Lambda}_m, \kappa_n+\tilde{\kappa}_m) = \frac{1}{\Gamma(n+m-\kappa_n-\tilde{\kappa}_m)}(\Lambda_n+\tilde{\Lambda}_m)^{n+m-\kappa_n-\tilde{\kappa}_m}\lambda^{n+m-\kappa_n-\tilde{\kappa}_m-1}e^{-(\Lambda_n+\tilde{\Lambda}_m)\lambda}. \tag{76}$$

Thus, we know that the average relative entropy, i.e. conditional mutual information given the historical data $\mathbf{g}^{\mathbf{n}}$ and $\mathbf{z}^{\mathbf{n}}$:

$$\begin{aligned}&\mathbf{I}(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}}; \theta|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}) \\ &\quad= \mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}}, \mathbf{r}^{\mathbf{n}})}\mathbf{D}_{KL}(\pi_{\mathbb{P}}(p, \lambda|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}}, \tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}})||\pi_{\mathbb{P}}(p, \lambda|\mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})) \\ &\quad= \mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}}, \mathbf{r}^{\mathbf{n}})}\mathbf{D}_{KL}(\pi_{\mathbb{P}}(p|\kappa_n + \tilde{\kappa}_m)||\pi_{\mathbb{P}}(p|\kappa_n)) \\ &\qquad+ \mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}}, \mathbf{r}^{\mathbf{n}})}\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\lambda|\Lambda_n + \tilde{\Lambda}_m, \kappa_n + \tilde{\kappa}_m)||\pi_{\mathbb{P}}(\lambda|\Lambda_n, \kappa_n))\end{aligned} \tag{77}$$

We are going to employ the following two simple facts regarding relative entropy between two Beta distributions and two Gamma distributions, respectively. Denote $\psi(x) := \dot{\Gamma}(x)/\Gamma(x)$ to be the Digamma function. Suppose $f(x;\alpha,\beta)$ is the probability density function for Beta distribution $\mathbf{Beta}(\alpha,\beta)$, then

$$\begin{aligned}&\mathbf{D}_{KL}(f(x;\alpha_0,\beta_0)||f(x;\alpha_1,\beta_1))\\&\quad=\ln\left[\frac{B(\alpha_1,\beta_1)}{B(\alpha_0,\beta_0)}\right]+(\alpha_0-\alpha_1)\psi(\alpha_0)+(\beta_0-\beta_1)\psi(\beta_0)+(\alpha_1-\alpha_0+\beta_1-\beta_0)\psi(\alpha_0+\beta_0)\end{aligned}\tag{78}$$

Suppose $g(x;\alpha,\beta)$ is the density function for Gamma distribution $\mathbf{Gamma}(\alpha,\beta)$, then

$$\begin{aligned}&\mathbf{D}_{KL}(g(x;\alpha_0,\beta_0)||g(x;\alpha_1,\beta_1))\\&\quad=(\alpha_0-\alpha_1)\psi(\alpha_0)-\ln\Gamma(\alpha_0)+\ln\Gamma(\alpha_1)+\alpha_1(\ln\beta_0-\ln\beta_1)+\alpha_0\frac{\beta_1-\beta_0}{\beta_0}.\end{aligned}\tag{79}$$

Applying (78) we know that

$$\begin{aligned}&\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\mathbf{D}_{KL}(\pi_{\mathbb{P}}(p|\kappa_n+\tilde{\kappa}_m)||\pi_{\mathbb{P}}(p|\kappa_n))\\&\quad=\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[\ln\left(\frac{B(1/2+n-\kappa_n,1/2+\kappa_n)}{B(1/2+n+m-\kappa_n-\tilde{\kappa}_n,1/2+\kappa_n+\tilde{\kappa}_n)}\right)\right]\\&\qquad+\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[(m-\tilde{\kappa}_m)\psi(1/2+n+m-\kappa_n-\tilde{\kappa}_m)\right]\\&\qquad+\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[\tilde{\kappa}_m\psi(1/2+\tilde{\kappa}_m+\kappa_n)\right]\\&\qquad-m\psi(1+n+m)\end{aligned}\tag{80}$$

Applying (79) we know that

$$\begin{aligned}&\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\lambda|\Lambda_n+\tilde{\Lambda}_m,\kappa_n+\tilde{\kappa}_m)||\pi_{\mathbb{P}}(\lambda|\Lambda_n,\kappa_n))\\&\quad=\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[(m-\tilde{\kappa}_m)\psi(n+m-\kappa_n-\tilde{\kappa}_m)\right]\\&\qquad-\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[\ln\Gamma(n+m-\kappa_n-\tilde{\kappa}_m)-\ln\Gamma(n-\kappa_n)\right]\\&\qquad+\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[(n-\kappa_n)\ln\frac{\Lambda_n+\tilde{\Lambda}_m}{\Lambda_n}\right]\\&\qquad-\mathbb{E}^{(\tilde{\mathbf{g}}^{\mathbf{m}},\tilde{\mathbf{z}}^{\mathbf{m}})|(\mathbf{g}^{\mathbf{n}},\mathbf{r}^{\mathbf{n}})}\left[(n+m-\kappa_n-\tilde{\kappa}_m)\frac{\tilde{\Lambda}_m}{\Lambda_n+\tilde{\Lambda}_m}\right]\end{aligned}\tag{81}$$

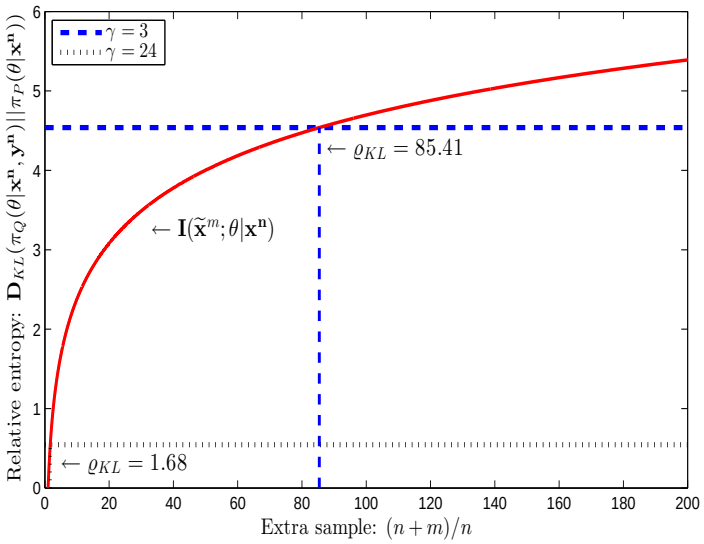


Figure 9: Computing the Finite Sample Information Ratio

We plug (80) and (81) back into (77), we can get the formula for the average relative entropy $\mathbf{I}(\tilde{\mathbf{g}}^{\mathbf{m}}, \tilde{\mathbf{z}}^{\mathbf{m}}; \theta | \mathbf{g}^{\mathbf{n}}, \mathbf{z}^{\mathbf{n}})$.

Figure 9 illustrates the procedure for computing the finite sample information ratio.

C.2.2 ABC Method and Implementation

Given the special structure of our problem, we propose a tilted ABC method in the hope of boosting the speed of our simulation. The algorithm described here is for the case of joint estimation with the risk aversion coefficient γ. The algorithm can be adapted easily for the special case where the value of γ is fixed.

The posterior for the unconstrained model satisfies

$$\begin{aligned} \theta, \phi \mid \mathbf{r}, \mathbf{g}, \mathbf{z} \sim & \operatorname{Beta}\left(p | 0.5 + n - \kappa_n, 0.5 + \kappa_n\right) \\ & \otimes \left[\operatorname{IWishart}_{\nu_n}\left(\Sigma | S_n\right) \otimes \mathrm{N}((\mu, \eta) | \mu_n, \Sigma)\right] \\ & \otimes \operatorname{Gamma}\left(\lambda | n - \kappa_n, \sum_{t=1}^{n} z_t (g_t - \underline{v})\right) \\ & \otimes \chi^{-2}\left(\nu^2 | n - \kappa_n, s_n/(n - \kappa_n)\right), \end{aligned} \tag{82}$$

where

$$x_t = (g_t, r_t)^T, \quad \mu_n = \sum_{t=1}^{n}(1-z_t)x_t / \sum_{t=1}^{n}(1-z_t), \quad \kappa_n = \sum_{t=1}^{n}(1-z_t), \quad \nu_n = \kappa_n - 1,$$
$$S_n = \sum_{t=1}^{n}(1-z_t)(x_t-\mu_n)(x_t-\mu_n)^T, \quad s_n = \sum_{t=1}^{n} z_t(r_t - bg_t)^2.$$

Define

$$\overline{g} = \sum_{t=1}^{n}(1-z_t)g_t/\kappa_n \quad \text{and} \quad \overline{r} = \sum_{t=1}^{n}(1-z_t)r_t/\kappa_n.$$

The posterior of the constrained model satisfies:

$$\begin{aligned}
\pi_{\mathbb{Q}}(\theta, \phi | \mathbf{g^n}, \mathbf{r^n}, \mathbf{z^n}) \propto\ & p^{n-\kappa_n+1/2-1}(1-p)^{\kappa_n+1/2-1} \\
& \times |\Sigma|^{\nu_n/2} \exp\left(-\frac{1}{2} tr(S_n \Sigma^{-1})\right) \\
& \times \sqrt{\kappa_n}\sigma^{-1} \exp\left(-\frac{\kappa_n}{2\sigma^2}(\mu - \overline{g})^2\right) \\
& \times \tau^{-1}(1-\rho^2)^{-1/2} \times \exp\left(-\frac{\kappa_n}{2(1-\rho^2)\tau^2}\left(\eta(\theta,\phi) - \overline{r} - \rho\frac{\tau}{\sigma}(\mu - \overline{g})\right)^2\right) \\
& \times \mathbf{1}_{(\eta(\theta,\phi) > \underline{\eta}^*)} \\
& \times \mathbf{1}_{(\lambda > \gamma)} \lambda^{n-\kappa_n-1} \exp\left(-\lambda \sum_{t=1}^{n} z_t(-g_t - \underline{v})\right) \\
& \times \mathbf{1}_{(\nu > 0)} \frac{1}{\sqrt{2\pi}(\nu^2)^{3/2}} \exp\left\{-\frac{1}{2\nu^2}\sum_{t=1}^{n} z_t\left(r_t - bg_t\right)^2\right\} \\
& \times \mathbf{1}_{(1 \le \gamma \le 50)}
\end{aligned} \tag{83}$$

Then, the posterior distribution will not change if we view the model in a different way as follows:

$$\overline{r} \sim \mathrm{N}\left(\eta(\theta,\phi) + \rho\frac{\tau}{\sigma}(\overline{g} - \mu), \tau^2(1-\rho^2)\right) \quad \text{where} \quad \eta(\theta,\phi) > \underline{\eta}^*,$$

with priors

$$\gamma \sim \text{Uniform}[0, 50],$$
$$p \sim \text{Beta}(n - \kappa_n + 1/2, \kappa_n + 1/2),$$
$$\Sigma \sim \text{IWishart}_{\nu_n}(\Sigma | S_n),$$

$$\mu|\sigma^2 \sim \mathrm{N}(\overline{g}, \sigma^2/\kappa_n),$$
$$\lambda \sim \mathrm{Gamma}\left(\lambda|n-\kappa_n, \sum_{t=1}^{n} z_t(g_t - \underline{v}), \lambda > \gamma\right),$$
$$\nu^2 \sim \chi^{-2}\left(\nu^2|n-\kappa_n, s_n/(n-\kappa_n)\right).$$

The tilted ABC method is implemented as follows.

Algorithm We illustrate the algorithm for simulating samples from the posterior (83) based ABC method. We choose the threshold in ABC algorithm as $\epsilon = \hat{\tau}/n/100$, where $\hat{\tau}$ is the sample standard deviation of the observations $r_1, \cdots, r_n$. Our tilted ABC algorithm can be summarized as follows:

For step $i = 1, \cdots, N$:

Repeat the following simulations and calculations:

(1) simulate $\tilde{\gamma} \sim \mathrm{Uniform}[0, 50]$,

(2) simulate $\tilde{p} \sim \mathrm{Beta}(n - \kappa_n + 1/2, \kappa_n + 1/2)$,

(3) simulate $\tilde{\Sigma} \sim \mathrm{IWishart}_{\nu_n}(\Sigma|S_n)$,

(4) simulate $\tilde{\mu}|\tilde{\sigma}^2 \sim \mathrm{N}(\overline{g}, \sigma^2/\kappa_n)$,

(5) simulate $\tilde{\lambda} \sim \mathrm{Gamma}\left(\lambda|n-\kappa_n, \sum_{t=1}^{n} z_t(g_t - \underline{v}), \lambda > 0\right)$,

(6) simulate $\tilde{\nu}^2 \sim \chi^{-2}\left(\nu^2|n-\kappa_n, s_n/(n-\kappa_n)\right)$,

(7) calculate $\tilde{\eta} = \eta(\tilde{\theta}, \tilde{\phi})$ with

$$\tilde{\theta} = (\tilde{p}, \tilde{\lambda}, \tilde{\mu}, \tilde{\sigma})$$
$$\tilde{\phi} = (\tilde{\tau}, \tilde{\rho}, \tilde{\nu})$$

(8) simulate $\tilde{r} \sim \mathrm{N}\left(\tilde{\eta} + \tilde{\rho}\frac{\tilde{\tau}}{\tilde{\sigma}}(\overline{g} - \tilde{\mu}), \tilde{\tau}^2(1-\tilde{\rho}^2)\right)$,

Until (i) $|\tilde{r} - \overline{r}| < \epsilon$ and (ii) $\tilde{\eta} > \underline{\eta}^*$, we record

$$\theta^{(i)} = \tilde{\theta}$$
$$\phi^{(i)} = \tilde{\phi}$$

Set $i = i + 1$, if $i < N$; end the loop, if $i = N$.

Using this algorithm, we shall get simulated samples $\theta^{(1)}, \cdots, \theta^{(N)}$ from the posterior (83).