Measuring the "Dark Matter" in Asset Pricing Models

Hui Chen Winston Wei Dou Leonid Kogan*

February 23, 2017

Abstract

We propose a new quantitative measure of model fragility, based on the tendency of a model to over-fit the data in sample with poor out-of-sample performance. We formally show that structural economic models are fragile when the cross-equation restrictions they impose on the baseline model appear excessively informative about combinations of model parameters that are otherwise difficult to estimate. Our fragility measure is analytically tractable, which is helpful to identify the parameter combinations as sources of model fragility. Using these new tools, we diagnose fragility in asset pricing models with rare disasters and long-run consumption risk.

JEL Codes: C1, D83, E44, G12

Keywords: Cross-equation restriction; Disaster risk; Information theory; Kullback-Leibler divergence; Long-run risk; Model fragility; Robustness; Robust Bayesian

*Chen: MIT Sloan and NBER (huichen@mit.edu). Dou: Wharton (wdou@wharton.upenn.edu). Kogan: MIT Sloan and NBER (lkogan@mit.edu). We thank Bradyn Breon-Drish, Alon Brav, Murray Carlson, John Cochrane, George Constantinides, Ian Dew-Becker, Lars Peter Hansen, Karen Lewis, Christian Opp, Frank Schorfheide, Martin Schneider, Luke Taylor, Laura Veldkamp, Tan Wang, S. Viswanathan, Fernando Zapatero, Stan Zin, and seminar participants at CITE, INSEAD, ITAM Finance Conference, Macrofinance Workshop, MIT Sloan, Northwestern, Chicago, Harvard, NYU, NBER Asset Pricing and Capital Markets Meetings, Red Rock Finance Conference, UBC Winter Finance Conference, Utah Winter Finance Conference, WFA, WashU, Wharton, and Yale SOM for comments, and Ye Luo for his comments on the proofs. All errors are our own.

1 Introduction

When building and evaluating a quantitative economic model, we care about how the model performs out of sample in addition to how well it fits the past data. This dual concern gives rise to the classic tradeoff between the accuracy of in-sample fit and the tendency of over-fitting. Too much emphasis on in-sample fit favors complex models, which are prone to over-fit the data in sample and likely to have poor out-of-sample performance. Precisely, those complex structural models over-utilize degrees of freedom of some parameters to accommodate certain economic restrictions implied by structural components in obtaining accurate in-sample fit. We refer to such economic restrictions (structural components) as fragile and such parameters as “dark matter”. A model, containing such fragile structural components or “dark matter”, is also referred to as fragile. Model fragility is a property of a model which captures its tendency to over-fit the past data, or in other words, captures the unreliability to conclude its out-of-sample performance based on the accuracy in-sample fit. Thus, models with higher fragility should be less favored among a set of candidate models that fit the past data well.

The above tradeoff is intuitive but not easy to implement in practice. As we build increasingly sophisticated quantitative structural models, the need for a systematic way to quantify model fragility also grows. Traditional over-fitting tendency measures, including the Akaike Information Criterion (AIC) and its variants, focus on the number of free parameters in a model used to accommodate its functional forms. Such measures potentially miss the implicit complexity: the effective degrees of freedom in a model depends not only on the number of free parameters, but also on the sensitivity of the key model implications to “reasonable” perturbations of parameter values. If its implications are highly sensitive, a particular economic restriction of the model can always fit the data by choosing specific parameter values, and thus it tends to over-fit the data in sample. In such case, the accuracy of in-sample fit becomes unreliable for assessing the particular structural components and, of course, the full model.

In this paper, we propose a new quantitative measure of model fragility from an informational perspective. Our measure is constructed based on Fisher information matrices, so we refer to it as Fisher fragility measure. Consider a typical structural model as a combination of functional-form specifications and parameters implied by economic theories and statistical distributions. The model describes a joint distribution of variables $\mathbf{x}_t$ and $\mathbf{y}_t$.[1] The baseline model describes the distribution of sample $\mathbf{x^n} \equiv (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ using the parameter vector $\theta \in \Theta$. The structural component assumed on top of the baseline model is chosen to be evaluated for its fragility. We are therefore measuring the "dark matter" of parameters θ of the baseline model. The structural component, on top of the baseline model, introduces additional ingredients that establish a joint distribution of $(\mathbf{x^n}, \mathbf{y^n})$.[2] To be more precise, we think of the additional economic restrictions implied by economic theories as adding *cross-equation restrictions* to the system of moments based solely on the baseline model. By definition, our Fisher fragility measure effectively compares the inverse Fisher information matrices for the baseline model and the full structural model along the directions associated with these linear subspaces and aggregates the differences.

Our Fisher fragility measure provides a simple decomposition that attributes the sources of model fragility (i.e. "dark matter") to a set of 1-dimensional linear subspaces of the parameter space. This decomposition offers an intuitive sample size interpretation. Each 1-dimensional linear subspace, indexed by j, corresponds to a particular linear combination of model parameters, $v_j\theta$. We assume that model parameters are estimated by GMM, with the fitted parameter vector $\widehat{\theta}$, and use the GMM J-distance, $J(\widehat{\theta}; \mathbf{x^n}, \mathbf{y^n})$, as the quantitative measure of model's in-sample fit given observations $(\mathbf{x^n}, \mathbf{y^n})$. Asymptotically, our measure corresponds to the amount of extra data needed to lower the asymptotic variance of $v_j\widehat{\theta}$ under the baseline model to the level of its variance under the full structural model (with the original data).

[1] We explain the details of the generic theoretical framework in Section 2.

[2] As an example, consider a Lucas-tree economy. There is a representative agent with certain preferences, and the growth rate of endowment is IID normal. A structural model in this case can be the model for the joint dynamics of the exogenous endowment ($\mathbf{x}_t$) and endogenous return on the endowment claim ($\mathbf{y}_t$), and θ includes the mean and volatility of endowment growth.

It is worth highlighting that when measuring model fragility, the goal is not to prove a model wrong. It is true that if a model is misspecified, further testable restrictions may reveal that. However, as Box (1976) and Hansen (2014) stress, all models are simplifications of reality that can eventually be rejected with sufficient data. Hansen (2014) states that "the important criticisms are whether our models are wrong in having missed something essential to the questions under consideration." This is also why we formulate our measure using the GMM framework. Through the selection of moment conditions, the econometrician has the ability to determine what the essential predictions of the model are.

It is also worth emphasizing that when measuring model fragility, the goal is not to estimating parameters. This paper is not really proposing new estimation procedures or drawing statistical inferences of any point estimators. Rather, the main purpose is to provide a new model fragility measure facilitating structural model selection when there are multiple candidates that fit a common set of fixed observations well in sample. Our model fragility measure is in the same spirit of those penalization procedures based on statistical fragility measures and adopted in statistical model selections such as AIC, BIC, and LASSO procedures. However, differently, our measure is specifically constructed for structural economic models. The over-fitting (model fragility) evaluation is opposite to the goodness-of-fit consideration; the latter takes a parametric model as fixed and focuses on the distribution of possible sample generated from it, whereas the former takes a sample as given and focuses on the sensitivity of various parametric models that fit in sample.

How to justify that our measure is indeed a measure of model fragility? To answer the important question, we extend a popular measure of statistical over-fitting tendency to our structural setting,[3]and we establish an asymptotic equivalence

[3]Spiegelhalter, Best, Carlin, and van der Linde (2002) measures over-fitting in a similar way, but using the log-likelihood instead of the J-distance. Also, they use arbitrary prior for generating the "reasonable" alternative models to assess a statistical model's over-fitting tendency; however, we argue it's crucial to choose a baseline model and a self-coherent posterior for generating the "reasonable" alternative functional-form specifications of structural components in economic modeling evaluation. This procedure allows for more economic-meaningful model assessment, beyond pure

result showing that the over-fitting tendency measure is actually equivalent to our Fisher fragility measure. Let θ_0 denote the parameter value for the true model. The corresponding value of the J-distance, $J(\theta_0; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}})$ is generally higher than the fitted value $J(\widehat{\theta}(\mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}); \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}})$, because the latter is chosen to minimize the J-statistic in sample. Then, the gap between the J-distance for the true model and the fitted structural model, $d\{\theta_0; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\} = J(\theta_0; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}) - J(\widehat{\theta}(\mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}); \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}})$, measures the degree of over-fitting by the estimated model. Not knowing what the true model is, we follow the common approach and average the degree of over-fitting over a set of possible true models, $\int_{\theta \in \Theta} \xi(\theta) d\{\theta; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\}\, d\theta$, where $\xi(\theta)$ assigns relative weights to alternative models. There is no broadly accepted choice of how to weigh the alternative models, and the exact specification depends on the context. We first specify the baseline model for $\mathbf{x}_t$. We then use the posterior distribution for θ implied by the baseline model, $\pi(\theta|\mathbf{x}^{\mathbf{n}})$, as the distribution over the alternative models $\xi(\theta)$. With this definition, we are assuming that inference based on the baseline model is reliable. We are therefore measuring the fragility of the full structural model relative to the baseline model.

Sensitivity analysis is a common technique for assessing model robustness. Intuitively, a model is considered robust if its key implications are not excessively sensitive to small perturbations of model parameters. In practical applications, one must specify the relevant perturbations and quantify "excessive sensitivity." As a result, it is difficult to generalize the traditional sensitivity analysis to multivariate settings. Model fragility may not be fully revealed by perturbing individual parameters – one must contemplate all possible multivariate perturbations, making the common approach impractical for high-dimensional problems.

Our methodology is not subject to such limitations. We use the posterior associated with the baseline model to weigh the relevant perturbations, and use the variance of the moments in the structural model to judge the degree of sensitivity of the moments.

statistical considerations. Using this procedure, economists can focus on the fragility of certain economic restrictions implied by economic theories and the "dark matter" of certain parameter space, not necessarily the whole model and all its parameters.

This eliminates the need for ad hoc choices associated with traditional sensitivity analysis. In addition, the asymptotic measure helps diagnose the sources of model fragility. Knowing the relative importance of each subspace for the overall fragility of the model effectively reduces the dimensionality of the multivariate sensitivity analysis. For example, if a single 1-dimensional subspace is dominant in terms of its contribution to overall model fragility, one only needs to examine the sensitivity of various moments to the perturbation of parameters in this particular subspace to quantify the main aspects of model fragility.

As a theoretical contribution, we connect the model fragility measure to the informativeness of the economic restrictions on the model parameters. To introduce the concept of informativeness, consider an example of a model that links the observations of the stock price P to the parameter θ describing the distribution of cash flows through a restriction: $\mathbb{E}[P] = \bar{P}(\theta)$. An econometrician starts with a baseline statistical model for cash flows and forms an (unconstrained) posterior belief about θ based on the observed cash-flow data and the baseline model, which is depicted in the left panel of Figure 1. The flatness of this posterior distribution indicates that there is nontrivial uncertainty about the true value of θ according to the baseline model.

The middle panel plots the model-implied price function $\bar{P}(\theta)$. Due to the high sensitivity of $\bar{P}$ to θ (the derivative $\partial\bar{P}/\partial\theta$ is large), there is only a narrow set of values of θ (highlighted by the shaded region) for which the observed price data are statistically close to the model-implied prices. This has two implications. First, by imposing the economic model, the econometrician obtains a posterior for θ (see the right panel) that is much more concentrated than the posterior distribution under the baseline model. In this case, we say that the economic restriction $\mathbb{E}[P] = \bar{P}(\theta)$ is highly informative about θ.

Second, for values of θ away from the shaded region in the middle panel but still in the range of values considered highly likely under the unconstrained posterior, the fit between the model and the observed price data deteriorates drastically, which is a sign of model fragility. Thus, high informativeness of the economic restrictions is

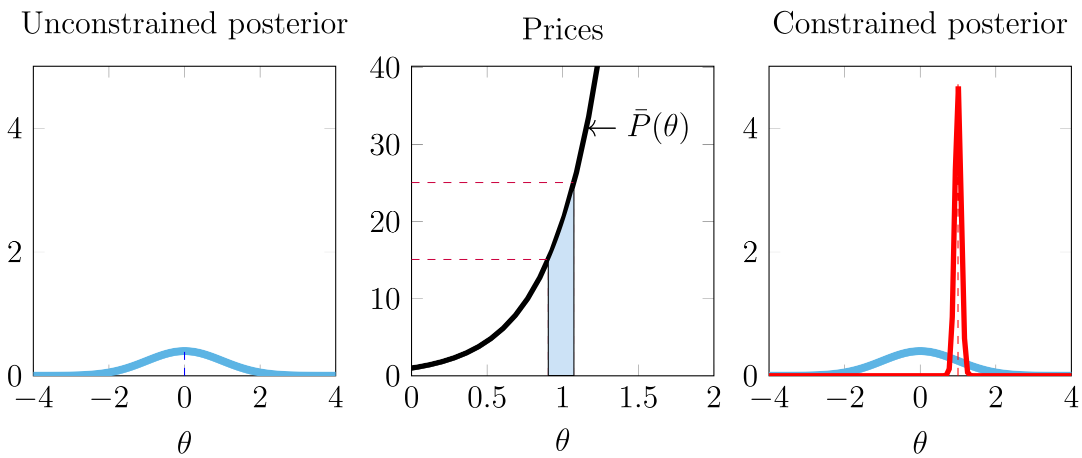

Figure 1: An example of an "informative" economic restriction on the parameters. The left panel plots the unconstrained posterior about θ based on cash-flow data. The middle panel plots the price function $\bar{P}(\theta)$. The dashed lines represent the confidence band for the mean of price observations. The right panel plots the constrained posterior about θ based on both cash-flow and price data.

closely linked to model fragility. Economic parameter restrictions are highly informative when they can significantly influence inference about certain combinations of model parameters that are relatively difficult to estimate statistically without such restrictions. Such parameter combinations are where the "dark matter" concentrates in the parameter space.

We formalize the above notion of informativeness of economic restrictions in an information-theoretic framework with an intuitive effective sample size interpretation. The informativeness of cross-equation restrictions relative to the baseline model is also reflected in the effect of the former on the posterior distribution of model parameters. We quantify the discrepancy between the posteriors of model parameters under the baseline model and under the model with further economic restrictions using relative entropy. We then define an effective sample size measure of informativeness of cross-equation restrictions as the average amount of extra data that, under the baseline model, generates the same magnitude of the shift in the posterior distribution. In other words, we equate the information content of the economic restrictions with the information content of additional data under the baseline statistical model. We show that the resulting measure of informativeness of cross-equation restrictions is related asymptotically to our measure of model fragility.

An important class of applications of our measure is to structural models that involve agents' beliefs. One prevalent approach to discipline beliefs is by imposing the rational expectations (RE) assumption. The RE assumption ties down the beliefs of economic agents by endowing them with precise knowledge of the probability law implied by an economic model. A common example of a RE model is a setting in which the agents know the true parameter value θ_0. The assumption of such precise knowledge is usually justified as a limiting approximation to the beliefs formed by learning from a sufficiently long history of data (see Hansen, 2007). If the posterior distribution of θ given $\mathbf{x^n}$ under the baseline model serves to describe the outcome of such learning, then high fragility of the economic model means that the model moments are highly sensitive to the exact choice among the likely values of the model parameter vector. In that case, assuming that the agents know the true parameter vector may be a poor approximation to a broader class of models in which agents maintain nontrivial uncertainty about the probability law of the model.

We apply the fragility measure to two examples from the asset pricing literature. The first example is a rare-disaster model. In this model, parameters describing the likelihood and the magnitude of economic disasters are relatively difficult to estimate from the data unless one uses information in asset prices.[4] We describe the fragility measure in this example analytically. We also illustrate how to incorporate

[4]A few papers have pointed out the challenges in testing disaster risk models. Zin (2002) shows that certain specifications of higher-order moments in the endowment growth distribution can help the model fit the empirical evidence while being difficult to reject in the data. In his 2008 Princeton Finance Lectures, John Campbell suggests that variable risk of rare disasters might be the "dark matter for economists." A related concept is the so-called "Peso problem" (see, e.g., Rogoff, 1980; Krasker, 1980; Veronesi, 2004; Lewis, 2008). The effect focuses on forecast errors that are systematically mistaken ex post in finite sample when agents' expectations about infrequent discrete shifts do not materialize in their observations. Thus, the term is interchangeable with the finite-sample weak identification problems. In contrast, the concept "dark matter" differs in three aspects: (1) it is a model property but not a finite-sample issues; (2) it can be huge even when parameters are well-identified in finite sample or the other way around; (3) it is not only about dynamics of rare discrete events, but also about preference or learning specifications when the baseline is structural.

uncertainty about the structural parameters (preference parameters in this context) when computing model fragility. The second example is a long-run risk model with a six-dimensional parameter space. We use this example to illustrate how to systematically diagnose the sources of fragility in a complex model.

1.1 Related Literature

The idea that model fragility is connected to complexity dates back at least to Fisher (1922). Model complexity is traditionally measured by the number of parameters in the model, because of the coincidence of the two quantities in Gaussian-linear models (see, e.g. Ye, 1998; Efron, 2004). Numerous statistical model selection procedures are based on this idea.[5]

The limitations of using the number of parameters to measure model complexity are well known. Extant literature covers several alternative approaches to measuring the "implicit model complexity." Ye (1998), Shen and Ye (2002), and Efron (2004) propose to measure complexity (or "generalized degrees of freedom" in their terminology) for Gaussian-linear models using the sensitivity of fitted values with respect to the observed data. Gentzkow and Shapiro (2013) apply a similar idea to examine identification issues in complex structural models. Spiegelhalter, Best, Carlin, and van der Linde (2002), Ando (2007) and Gelman, Hwang, and Vehtari (2013), among others, propose a Bayesian complexity measure they call "the effective number of parameters," which is based on out-of-sample model performance. These methods measure the sensitivity of model implications to parameter perturbations. The important common feature of these proposals is that they rely on the model being evaluated to determine the magnitude of necessary parameter perturbations. This is potentially problematic when evaluating economic models that are fragile according to our definition. For such models, the posterior distribution over the parameters is highly concentrated as a result

[5]Examples include the Akaike Information Criterion (AIC) (Akaike, 1973), the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Risk Inflation Criterion (RIC) (Foster and George, 1994), and the Covariance Inflation Criterion (CIC) (Tibshirani and Knight, 1999).

of excessive model sensitivity to its parameters. Relying on this posterior to generate parameter perturbations can under-represent the true extent of model fragility. In contrast, we propose to use the baseline model to determine the distribution $\xi(\theta)$ over the potential alternative models.

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.[6] In particular, the literature on robustness in macroeconomic models (see Hansen and Sargent, 2008; Epstein and Schneider, 2010, for recent surveys) recognizes that the traditional assumption of agents' precise knowledge of the relevant probability distributions is not reasonable in certain contexts. This literature explicitly incorporates 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 fragility of economic models, thus identifying situations in which parameter uncertainty and robustness could be particularly important.

Our work is connected to the literature in rational expectations econometrics, where economic assumptions (the cross-equation restrictions) have been used extensively to gain efficiency in estimating the structural parameters.[7] When imposing such assumptions results in a fragile model, standard inference may result in excessively small confidence regions for the parameters, with low coverage probability under reasonable parameter perturbations. Related, fragile models tend to generate excessively high quality of in-sample fit, which biases model selection in their favor. The combination of these two effects makes common practice of post-selection inference misleading in the presence of "dark matter".

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

[7]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).

2 Measuring Model Fragility

In this section, we first introduce a formal measure of model fragility. Then we derive asymptotic properties of the fragility measure.

2.1 A Generic Model Structure

Consider a baseline model $\mathcal{P}$, which is a part of the full structural model $\mathcal{Q}$. The baseline model $\mathcal{P}$ specifies the dynamics of a vector of variables $\mathbf{x_t}$ with the underlying distribution $\mathbb{P}$. In comparison, the full structural model $\mathcal{Q}$ aims to capture certain economic features of the distribution $\mathbb{Q}$ that governs the joint dynamics of $\mathbf{x_t}$ and additional variables $\mathbf{y_t}$.

More precisely, the baseline model $\mathcal{P}$ of $\mathbf{x^n}$ can be specified by a $D_\Theta \times 1$ parameter vector θ, while the whole structural model $\mathcal{Q}$ may incorporate extra data $\mathbf{y^n}$ and parameters ψ. The distribution of $\mathbf{y^n}$ conditional on $\mathbf{x^n}$ depends on not only the baseline parameter vector θ, but also the $D_\Psi \times 1$ nuisance parameter vector ψ. The baseline parameters θ are over-identified by both $\mathbb{P}$ and additional structural components in $\mathcal{Q}$. The nuisance parameters ψ are parameters which are not part of the baseline model but should be accounted for in assessing the full model. We assume that the true parameter values θ_0 and ψ_0 are contained in the interiors of sets Θ and Ψ, respectively.

We assume that the stochastic process $\{\mathbf{x_t}\}$ is strictly stationary and ergodic with a stationary distribution $\mathbb{P}$. The true joint distribution for $\mathbf{x^n} \equiv (\mathbf{x_1}, \cdots, \mathbf{x_n})$ is $\mathbb{P}_n$. Similarly, we assume that the joint stochastic process $\{\mathbf{x_t}, \mathbf{y_t}\}$ is strictly stationary and ergodic with a stationary distribution $\mathbb{Q}$. The econometrician does not need to specify the full functional form of the joint distribution of $(\mathbf{x^n}, \mathbf{y^n}) \equiv \{(\mathbf{x_t}, \mathbf{y_t}) : \ t = 1, \cdots, n\}$, which we denote by $\mathbb{Q}_n$. The unknown joint density is $q(\mathbf{x^n}, \mathbf{y^n})$.

We evaluate the performance of a structural model under the Generalized Method of Moments (GMM) framework. The seminal paper by Hansen and Singleton (1982) pioneers the literature of applying GMM to evaluate rational expectation asset pricing models. Specifically, we assume that the model builder is concerned with the model's in-sample and out-of-sample performances as represented by a set of moment conditions,[8] based on a $D_{\mathbb{Q}} \times 1$ vector of functions $g_{\mathbb{Q}}(\theta, \psi; \mathbf{x}, \mathbf{y})$ of data observations $(\mathbf{x}_t, \mathbf{y}_t)$ and the parameter vectors θ and ψ satisfying the following conditions:

$$\mathbb{E}\left[g_{\mathbb{Q}}(\theta_0, \psi_0; \mathbf{x_t}, \mathbf{y_t})\right] = 0. \tag{1}$$

The baseline moment functions $g_{\mathbb{P}}(\theta; \mathbf{x_t})$ characterize the moment conditions of the baseline model. They constitute the first $D_{\mathbb{P}}$ elements of the whole vector of moment functions $g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t})$. Thus, the baseline moments can be represented by the full set of moments weighted by a special matrix:

$$g_{\mathbb{P}}(\theta; \mathbf{x_t}) = \Gamma_{\mathbb{P}} g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t}) \quad \text{where} \quad \Gamma_{\mathbb{P}} \equiv \left[I_{D_{\mathbb{P}}}, O_{D_{\mathbb{P}} \times (D_{\mathbb{Q}} - D_{\mathbb{P}})}\right]. \tag{2}$$

The moment functions $g_{\mathbb{P}}(\theta; \mathbf{x_t})$ depend only on parameters θ, since all parameters of the baseline model are included in θ. Accordingly, the moment conditions for the baseline model is

$$\mathbb{E}\left[g_{\mathbb{P}}(\theta_0; \mathbf{x_t})\right] = 0. \tag{3}$$

Denote the empirical moment conditions for the full model and the baseline model by

$$\widehat{g}_{\mathbb{Q},n}(\theta, \psi) \equiv \frac{1}{n} \sum_{t=1}^{n} g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t}) \text{ and } \widehat{g}_{\mathbb{P},n}(\theta) \equiv \frac{1}{n} \sum_{t=1}^{n} g_{\mathbb{P}}(\theta; \mathbf{x_t}), \text{ respectively.}$$

Then, the optimal GMM estimator $(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}})$ of the full model and that of the baseline

[8] We can also adopt the CUE method of Hansen, Heaton, and Yaron (1996) or its modification Hausman, Lewis, Menzel, and Newey (2011)'s RCUE method, or some other extension of GMM with the same first-order efficiency and possibly superior higher-order asymptotic properties. This will lead to alternative but conceptually similar measures of overfitting. To simplify the comparison with the Fisher fragility measure, we chose to use the original GMM framework.

model $\hat{\theta}^{\mathbb{P}}$ minimize, respectively,

$$\widehat{J}_{n,S_{\mathbb{Q}}}(\theta,\psi) \equiv n\widehat{g}_{\mathbb{Q},n}(\theta,\psi)^T S_{\mathbb{Q}}^{-1}\widehat{g}_{\mathbb{Q},n}(\theta,\psi) \quad \text{and} \quad \widehat{J}_{n,S_{\mathbb{P}}}(\theta) \quad \equiv n\widehat{g}_{\mathbb{P},n}(\theta)^T S_{\mathbb{P}}^{-1}\widehat{g}_{\mathbb{P},n}(\theta). \quad (4)$$

Here, $\widehat{J}_{n,S_{\mathbb{Q}}}$ and $\widehat{J}_{n,S_{\mathbb{P}}}$ are often referred to as the J-distances, and $S_{\mathbb{Q}}$ and $S_{\mathbb{P}}$ have the following explicit formulae (see Hansen, 1982),

$$S_{\mathbb{Q}} \equiv \sum_{\ell=-\infty}^{+\infty} \mathbb{E}\left[g_{\mathbb{Q}}(\theta_0,\psi_0;\mathbf{x_t},\mathbf{y_t})g_{\mathbb{Q}}(\theta_0,\psi_0;\mathbf{x_{t-\ell}},\mathbf{y_{t-\ell}})^T\right], \quad \text{and} \quad (5)$$

$$S_{\mathbb{P}} \equiv \sum_{\ell=-\infty}^{+\infty} \mathbb{E}\left[g_{\mathbb{P}}(\theta_0;\mathbf{x_t})g_{\mathbb{P}}(\theta_0;\mathbf{x_{t-\ell}})^T\right], \quad \text{respectively.} \quad (6)$$

The matrix $S_{\mathbb{Q}}$ and $S_{\mathbb{P}}$ are the covariance matrices of the moment conditions at the true parameter values. In practice, when $S_{\mathbb{Q}}$ or $S_{\mathbb{P}}$ is unknown, we can replace it with a consistent estimator $\widehat{S}_{\mathbb{Q},n}$ or $\widehat{S}_{\mathbb{P},n}$, respectively. The consistent estimators of the covariance matrices are provided by Newey and West (1987a), Andrews (1991), and Andrews and Monahan (1992).

We use GMM to evaluate model performance because of the concern of likelihood mis-specification. The GMM approach gives the model builder flexibility to choose which aspects of the model to emphasize when estimating model parameters and evaluating model specifications. This is in contrast to the likelihood approach, which relies on the full probability distribution implied by the structural model.

Finally, we introduce GMM Fisher information matrices. We denote the GMM Fisher information matrix for the baseline model as $\mathbf{I}_{\mathbb{P}}(\theta)$ (see Hansen, 1982; Hahn, Newey, and Smith, 2011), and

$$\mathbf{I}_{\mathbb{P}}(\theta) \equiv G_{\mathbb{P}}(\theta)^T S_{\mathbb{P}}^{-1} G_{\mathbb{P}}(\theta), \quad (7)$$

where $G_{\mathbb{P}}(\theta) \equiv \mathbb{E}\left[\nabla g_{\mathbb{P}}(\theta;\mathbf{x_t})\right]$,and for brevity, we denote $G_{\mathbb{P}} \equiv G_{\mathbb{P}}(\theta_0)$. We denote the analog for the structural model as $\mathbf{I}_{\mathbb{Q}}(\theta, \psi)$,

$$\mathbf{I}_{\mathbb{Q}}(\theta, \psi) \equiv G_{\mathbb{Q}}(\theta, \psi)^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}(\theta, \psi), \tag{8}$$

where $G_{\mathbb{Q}}(\theta, \psi) \equiv \mathbb{E}\left[\nabla g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t})\right]$, and for brevity, we denote $G_{\mathbb{Q}} \equiv G_{\mathbb{Q}}(\theta_0, \psi_0)$. Computing the expectation $G_{\mathbb{Q}}(\theta)$ and $G_{\mathbb{Q}}(\theta, \psi)$ requires knowing the distribution $\mathbb{Q}$. In cases when $\mathbb{Q}$ is unknown, $G_{\mathbb{P}}(\theta)$ in (7) and $G_{\mathbb{Q}}(\theta, \psi)$ in (8) can be replaced by their consistent estimators $\nabla g_{\mathbb{P}}(\theta; \mathbf{x_t})$ and $\nabla g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t})$. For the full model $\mathcal{Q}$, we will focus on its implied Fisher information matrix $\mathbf{I}_{\mathbb{Q}}(\theta|\psi)$:

$$\mathbf{I}_{\mathbb{Q}}(\theta|\psi) \equiv \left[\Gamma_\Theta \mathbf{I}_{\mathbb{Q}}(\theta, \psi)^{-1} \Gamma_\Theta^T\right]^{-1}, \quad \text{where } \Gamma_\Theta \equiv \left[I_{D_\Theta}, O_{D_\Theta \times D_\Psi}\right]. \tag{9}$$

More precisely, the Fisher information matrix $\mathbf{I}_{\mathbb{Q}}(\theta, \psi)$ can be partitioned into a two-by-two block matrix according to θ and ψ:

$$\mathbf{I}_{\mathbb{Q}}(\theta, \psi) = \begin{bmatrix} \mathbf{I}_{\mathbb{Q}}^{(1,1)}(\theta, \psi), & \mathbf{I}_{\mathbb{Q}}^{(1,2)}(\theta, \psi) \\ \mathbf{I}_{\mathbb{Q}}^{(2,1)}(\theta, \psi), & \mathbf{I}_{\mathbb{Q}}^{(2,2)}(\theta, \psi) \end{bmatrix}, \tag{10}$$

where $\mathbf{I}_{\mathbb{Q}}^{(1,1)}(\theta, \psi)$ is the $D_\Theta \times D_\Theta$ information matrix corresponding to baseline parameters θ, $\mathbf{I}_{\mathbb{Q}}^{(2,2)}(\theta, \psi)$ is the $D_\Psi \times D_\Psi$ information matrix corresponding to nuisance parameters ψ, and $\mathbf{I}_{\mathbb{Q}}^{(1,2)}(\theta, \psi) = \mathbf{I}_{\mathbb{Q}}^{(2,1)}(\theta, \psi)^T$ is the $D_\Theta \times D_\Psi$ cross-information matrix corresponding to θ and ψ. Then $\mathbf{I}_{\mathbb{Q}}(\theta|\psi)$ can be written as

$$\mathbf{I}_{\mathbb{Q}}(\theta|\psi) = \mathbf{I}_{\mathbb{Q}}^{(1,1)}(\theta, \psi) - \mathbf{I}_{\mathbb{Q}}^{(1,2)}(\theta, \psi) \mathbf{I}_{\mathbb{Q}}^{(2,2)}(\theta, \psi)^{-1} \mathbf{I}_{\mathbb{Q}}^{(1,2)}(\theta, \psi)^T, \tag{11}$$

which generally is not equal to the Fisher information sub-matrix $\mathbf{I}_{\mathbb{Q}}^{(1,1)}(\theta, \psi)$ for baseline parameters θ, except the special case in which $\mathbf{I}_{\mathbb{Q}}^{(1,2)}(\theta, \psi) = 0$, i.e. the knowledge of θ and that of ψ are not informative about each other. We assume that the information matrices are nonsingular in this paper (Assumption A4 in Appendix A).

By using the GMM Fisher information matrices, we appeal to the optimal weighting matrices $S_{\mathbb{P}}^{-1}$ and $S_{\mathbb{Q}}^{-1}$. In Hansen and Jagannathan (1997), it is shown that the optimal weighting matrix is not efficient for measuring model mis-specification. Hansen and Jagannathan (1997) try to measure model mis-specification but not the fragility of model's implications to potential mis-specifications (i.e. over-fitting tendency). They are very different objects. More precisely, to construct the lack-of-fit measure (i.e. measure of model mis-specification) in Hansen and Jagannathan (1997), the weighting matrix is intentionally chosen not to be the optimal GMM weighting matrix for the following reasons. First, the baseline metric for gauging the lack of fit should not be sensitive to the choice of SDF proxies under the assessment. This idea is similar to our idea of fixing a baseline model for over-fitting tendency measures. Second, the weighting matrix should not reward the sampling errors associated with the sample mean of the pricing errors. Since the absolute level of the pricing errors is a more reasonable measure for the mis-specification of a SDF proxy. In contrast, our over-fitting tendency measure requires model's goodness-of-fit measures as loss functions like in the statistical literature (see Spiegelhalter, Best, Carlin, and van der Linde, 2002). For the moment-based setting, as recommended by Hansen (1982), we adopt J-distance as the goodness-of-fit measure (loss function) which uses the efficient GMM weighting matrix. Following Hansen (1982), there have been further technical analysis that justify the unique role of optimal weighting matrix in gauging moment condition specifications from different angles. As emphasized by Newey and West (1987b), it is crucial to define the GMM likelihood ratio test statistics based on the optimal GMM estimator, for which $W = S_{\mathbb{Q}}^{-1}$. This is because another choice of the weighting matrix W will destroy the asymptotic property of having chi-squared distribution as limiting distribution and will break the asymptotic equivalence between GMM likelihood ratio test statistics and other GMM test statistics such as Wald and LM test statistics. Similarly, as highlighted in Kim (2002), it is important to use efficient weighting matrix for developing the theory of limited information likelihood based on (asymptotic) quadratic moment conditions.

2.2 Model Fragility

We now define our measure of model fragility. We start with introducing the "Fisher fragility measure" based on GMM Fisher information matrices and analyzing its properties.

Definition 1 (Fisher Fragility Measure)**.** *The Fisher fragility measure corresponding to a full-rank $D_{\mathbf{v}} \times D_\Theta$ matrix $\mathbf{v}$ is defined as*

$$\varrho^{\mathbf{v}}(\theta_0|\psi_0) \equiv \mathbf{tr}\left[\left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1}\left(\mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}^T\right)\right], \tag{12}$$

where $\mathbf{I}_{\mathbb{P}}(\theta_0)$ and $\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)$ are the matrices defined in (7) and (11), respectively.

In the special case where $\mathbf{v}$ is a full-rank $D_\Theta \times D_\Theta$ matrix, $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ is independent of the choice of $\mathbf{v}$. In that case we denote the Fisher fragility measure as $\varrho(\theta_0|\psi_0)$, which is the overall Fisher fragility measure of model. In another special case where $D_{\mathbf{v}} = 1$, our measure essentially becomes the ratio of two asymptotic variances of optimal GMM estimators $\mathbf{v}\hat{\theta}^{\mathbb{Q}}$ and $\mathbf{v}\hat{\theta}^{\mathbb{P}}$. From Definition 1, it immediately follows that the Fisher fragility measure can be characterized by the solution of an eigenvalue problem, which is summarized by Proposition 1 whose proof is in Appendix C.1.

Proposition 1. *For a full-rank $D_{\mathbf{v}} \times D_\Theta$ matrix $\mathbf{v}$, let $\lambda_1(\mathbf{v}) \geq \lambda_2(\mathbf{v}) \geq \cdots \geq \lambda_{D_{\mathbf{v}}}(\mathbf{v})$ be the eigenvalues of a Fisher-information-ratio matrix $\Pi_0(\mathbf{v})$ defined as follows*

$$\Pi_0(\mathbf{v}) \equiv \left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1/2}\left(\mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}^T\right)\left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1/2}. \tag{13}$$

Then the Fisher fragility measure can be characterized by the sum of the eigenvalues:

$$\varrho^{\mathbf{v}}(\theta_0|\psi_0) = \lambda_1(\mathbf{v}) + \lambda_2(\mathbf{v}) + \cdots + \lambda_{D_{\mathbf{v}}}(\mathbf{v}). \tag{14}$$

and the smallest eigenvalue $\lambda_{D_{\mathbf{v}}}(\mathbf{v})$ is not less than one.

The measure $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ is defined for specific feature directions $\mathbf{v}$ in the space of baseline parameters. One might be interested in searching among a class of feature directions to find the worst-case configuration. It is actually quite straightforward for the Fisher fragility measure, because the optimization problem is a finite-dimensional one. This leads us to define the following worst-case Fisher fragility measure.

Definition 2. *The worst-case Fisher fragility measure for the class of D-dimensional feature functions ($D \leq D_\Theta$) is defined as:*

$$\varrho^D(\theta_0|\psi_0) = \max_{\mathbf{v}\in\mathbb{R}^{D\times D_\Theta},\mathbf{Rank}(\mathbf{v})=D} \mathbf{tr}\left[\left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1}\left(\mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}^T\right)\right]. \tag{15}$$

The problem in (15) is a generalized eigenvalue problem. The following proposition summarizes its solution. The proof of Proposition 2 is in Appendix C.2.

Proposition 2. *Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{D_\Theta}$ be the eigenvalues of Fisher information ratio matrix $\Pi_0(I_{D_\Theta}) = \mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}$, with the corresponding $D_\Theta \times 1$ eigenvectors $e_1, e_2, \cdots, e_{D_\Theta}$. Then the D-dimensional worst-case Fisher fragility measure is equal to*

$$\varrho^D(\theta_0|\psi_0) = \lambda_1 + \lambda_2 + \cdots + \lambda_D, \tag{16}$$

with the worst-case D-dimensional linear subspace of the parameter space characterized by the matrix $\mathbf{v}_D^* = [v_1^*\ v_2^* \cdots v_D^*]^T$,

$$v_i^* = \mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}e_i \Big/ \left|\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}e_i\right|. \tag{17}$$

As a special case, the overall Fisher fragility measure is given by

$$\varrho(\theta_0|\psi_0) = \lambda_1 + \lambda_2 + \cdots + \lambda_{D_\Theta}. \tag{18}$$

From Proposition 2, it is easy to see that the worst-case Fisher fragility measure $\varrho^D(\theta_0|\psi_0)$ is monotonically increasing in the dimension of the subspace D.

Proposition 3. (Monotonicity) *For $D_1 \leq D_2 \leq D_\Theta$,*

$$\varrho^{D_1}(\theta_0|\psi_0) \leq \varrho^{D_2}(\theta_0|\psi_0). \tag{19}$$

The proof can be found in Appendix C.3.

2.3 An Informational Interpretation

The intuition behind the worst-case Fisher fragility measure is as follows. Through the matrix $\mathbf{v}$, we search over all D-dimensional linear subspaces of the parameter space to find the maximum discrepancy between the inverses of the two information matrices, $\mathbf{I}_{\mathbb{P}}(\theta_0)$ and $\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)$. In the context of GMM estimation (or other moment-based estimation), the inverse of the information matrix is linked to the asymptotic covariance matrices of the estimators. Since we require baseline model $\mathcal{P}$'s moment conditions to be included in those for the full model $\mathcal{Q}$, the asymptotic efficiency of the GMM estimator for the structural model dominates that of the baseline model. The Fisher fragility measure effectively compares the asymptotic covariance matrices of these two estimators to isolate the information provided by the structural model restrictions.

We can view Proposition 2 as a decomposition of the overall fragility of a model into D_Θ linear subspaces of 1 dimension. The i-th largest eigenvalue λ_i $(1 \leq i \leq D_\Theta)$ of $\Pi_0(I_{D_\Theta})$ gives the marginal contribution of the 1-dimensional linear subspace associated with v_i^* to the overall fragility measure. In the language of sensitivity analysis, such a decomposition reveals the directions in which small perturbations of parameters can have the largest impact on the model output. Moreover, the decomposition is also useful to gauge the tendency of obtaining extreme over-fitting outcomes for a structural model (see Corollary 2).

The Fisher fragility measure has a natural "effective-sample-size" interpretation. Consider the case of $D = 1$. In this case, we ask what is the minimum sample size required for the estimator of the baseline model to match or exceed the precision of the estimator for the full structural model in all 1-dimensional linear subspaces of the parameter space. Because the asymptotic covariance of the estimator is proportional to the sample size n, the required additional effective sample size for the baseline model is $\varrho^1(\theta_0|\psi_0)$ times the sample size n for the structural model to achieve at least the same estimation accuracy. The idea is formalized under a rigorous information-theoretic framework with the Kullback-Leibler divergence (i.e. the relative entropy) and finite-sample validity in the supplemental material Chen, Dou, and Kogan (2017). Therefore, the excessive informativeness of cross-equation restrictions is fundamentally associated with model fragility. More precisely, the fundamental idea of the Fisher fragility measure is that structural economic models are fragile when the cross-equation restrictions appear excessively informative about certain combinations of model parameters that are otherwise difficult to estimate ("dark matters").

2.4 Redundancy of Moment Conditions

It may appear that the Fisher fragility measure favors redundant cross-equation restrictions. To clarify that it is not the case, we emphasize that our work tries to answer different questions from the literature of efficient estimation and testing with correctly specified models. Our focus is related to efficient estimation with model uncertainty involves model selection. Under the setting of correct specifications, Breusch, Qian, Schmidt, and Wyhowski (1999) provide a general discussion of redundant moments, as well as the useful conditions in identifying redundant moment conditions in practice. Also, under the setting of correct specifications, Cheng and Liao (2015) propose a one-step procedure to distinguish valid-and-relevant moments from invalid or irrelevant moments and automatically achieve efficient point estimation with many moments. To be more precise, we focus on the setting in which multiple specifications are statistically valid in finite sample but there are over-fitting concerns in the statistical validity assessment. This in-sample over-fitting tendency is exactly what we are after, and it is an important issue since it leads to poor out-of-sample performance of models. Of course, if a model can be significantly rejected by efficient tests (i.e. statistical invalid), say by Hansen (1982) over-identification tests, there is no need to evaluate the fragility of that model.

3 Model Over-Fitting Tendency

The Fisher fragility measure, discussed in Section 2, is conceptually straightforward, and it is ready to be implemented in various applications. The intuitive interpretation of the measure is pure information-based in an asymptotic sense. How to justify that the measure is indeed a measure of model fragility? To address this question, we first introduce an econometric measure of over-fitting which extends a popular existing statistical over-fitting tendency measure to our structural setting (see Spiegelhalter, Best, Carlin, and van der Linde, 2002). We establish an asymptotic equivalence result showing that our Fisher fragility measure essentially quantifies the over-fitting tendency of certain structural component of model.

3.1 Econometric Measure of Over-fitting Tendency

Out theoretical results build on the Bayesian framework. The Bayesian framework is the most natural one to investigate model's over-fitting tendency (see Spiegelhalter, Best, Carlin, and van der Linde, 2002) and is a vastly used method for estimating macroeconomic structural models (see Herbst and Schorfheide, 2015). Let $\pi(\theta)$ be a prior distribution on θ, and let $\pi_{\mathbb{P}}(\theta|\mathbf{x}^{\mathbf{n}})$ be the posterior of θ based on the likelihood of the baseline model $\pi_{\mathbb{P}}(\mathbf{x}^{\mathbf{n}}|\theta)$. To establish the theoretical results involving Bayesian methods, we need to assume that the likelihood of the baseline model is well specified. The theoretical result in this section is the only place we assume full likelihood function for the baseline model. As far as the core idea is concerned, this assumption is not needed. When the likelihood function is well specified and the score functions are all used as moment restrictions in defining the Fisher fragility measure, the Fisher information matrix $\mathbf{I}_{\mathbb{P}}(\theta_0)$ becomes the standard one under the likelihood-based approach.

Definition 3 (Over-fitting Tendency Measure)**.** *We define the over-fitting tendency measure for the structural model $\mathcal{Q}$ relative to the baseline model $\mathcal{P}$ as*

$$\varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) \equiv \int d_{S_{\mathbb{Q}}}\{\theta; \mathbf{x^n}, \mathbf{y^n}\} \pi_{\mathbb{P}}(\theta|\mathbf{x^n}) d\theta, \tag{20}$$

$$\textit{where} \quad d_{S_{\mathbb{Q}}}\{\theta; \mathbf{x^n}, \mathbf{y^n}\} = \widehat{J}_{n,S_{\mathbb{Q}}}(\theta, \check{\psi}^{\mathbb{Q}}) - \widehat{J}_{n,S_{\mathbb{Q}}}(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}}). \tag{21}$$

Here, $\widehat{J}_{n,S_{\mathbb{Q}}}(\theta, \psi)$ is the J-distance defined in (4), and $(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}})$ is the GMM estimator, and $(\theta, \check{\psi}^{\mathbb{Q}})$ is the constrained GMM estimator with fixed θ. The covariance matrix $S_{\mathbb{Q}}$ theoretically depends on the true parameters θ_0 and ψ_0.

The idea of our fragility measure is to quantify the in-sample over-fitting of a structural model. In Equation (21), $d_{S_{\mathbb{Q}}}\{\theta; \mathbf{x^n}, \mathbf{y^n}\}$ is the J-distance (the GMM analog of the log likelihood ratio) of the model with jointly-fitted baseline parameter $\hat{\theta}^{\mathbb{Q}}$ and nuisance parameters $\hat{\psi}^{\mathbb{Q}}$, which provide the best in-sample fit of the data based on the GMM criterion, against an alternative model with baseline parameters θ and their fitted nuisance parameters $\check{\psi}^{\mathbb{Q}}$. Assuming true parameter is θ instead of $\hat{\theta}^{\mathbb{Q}}$, the fact that the J-distance based on $(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}})$ is smaller is a symptom of over-fitting.

The weights attached to alternative models are essential for our definition of model fragility. We consider alternative values of θ, while tuning the nuisance parameters ψ to fit the data as well as possible under the same criterion, i.e., we choose $\psi = \check{\psi}^{\mathbb{Q}}$, the constrained GMM estimator. Starting with a prior $\pi(\theta)$, we weigh the various alternative models using $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$ – the posterior for θ based on baseline model's moment conditions and data $\mathbf{x^n}$. The weighted average of $d_{S_{\mathbb{Q}}}\{\theta; \mathbf{x^n}, \mathbf{y^n}\}$ over the entire set of alternative models represents the tendency of model over-fitting.

The weights on various alternative models depend on $\mathcal{P}$, therefore $\varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ is a measure of over-fitting tendency of $\mathcal{Q}$ *relative to the baseline model* $\mathcal{P}$. That is, the measure $\varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ quantifies the degrees of freedom allowed by baseline parameters θ accommodating the additional structural restriction imposed by $\mathcal{Q}$ to achieve an accurate in-sample fit. Thus, the measure of model's over-fitting tendency (fragility) depends on the choice of the baseline model, or in other words, depends on the particular structural components of $\mathcal{Q}$ chosen to be assessed. For example, many structural models involve both a statistical model of exogenous variables and restrictions on the endogenous variables which are derived from the economic model. For such models, a natural choice may be to take the baseline $\mathbb{P}$ to be the statistical model, with $\mathbf{x_t}$ being the exogenous variables. Variables $\mathbf{y_t}$ would then be the endogenous ones in the structural model. In this context, our fragility measure quantifies the fragility of the structural model relative to the statistical model.

Alternatively, a structural model could be taken as the baseline model. Then, the fragility measure applies to the over-fitting caused by the additional economic restrictions imposed by $\mathcal{Q}$ relative to the baseline model $\mathcal{P}$.[9]

In general, there is no hard rule imposed on the choice of baseline model. These choices must be made by the model builder depending on which aspects of the model are intended to be covered by the fragility analysis. Thus, the choice of the baseline model, together with θ, is more economical and less statistical.

The distribution over the alternative models also depends on the choice of the prior $\pi(\theta)$. If the econometrician does not have any information about θ beyond the baseline model and the data $\mathbf{x^n}$, an "uninformative" prior would be a desirable choice, one candidate being the Jeffreys prior. In many cases a truly uninformative prior is difficult to define, especially in the presence of constraints. If the econometrician has additional information about θ outside the model (e.g., from additional data or other

[9]In this case, the preference parameters or belief parameters can be part of the baseline parameters θ. A natural case in which the preference parameters and belief parameters can show up as part of the baseline parameters θ is as follows. For example, consider a model that explains equity returns and equity options returns simultaneously. If the model of equity returns alone is taken as the baseline model, some preference parameters or belief parameters must be included in the baseline parameters θ. The goal is evaluate the fragility/informativeness of the cross-equation restrictions on the Euler equations of the equity options returns.

models), such information can be incorporated through an informative prior.

Our definition of model fragility builds upon Spiegelhalter, Best, Carlin, and van der Linde (2002), who propose a related measure of model complexity for statistical models. Our measure differs from theirs in two respects. First, we adopt the GMM framework as opposed to the likelihood framework to address the issue of stochastic singularities that arise in structural models and to give the econometrician the flexibility to focus on specific features of a model. Second, Spiegelhalter, Best, Carlin, and van der Linde (2002) do not explicitly specify the weighting of alternative models. By contrast, in economic modeling evaluation, it's crucial to choose a baseline model and use a self-coherent posterior to discipline the alternative structural component specifications. This procedure allows to focus on the fragility of certain economic restrictions implied by economic theories (i.e. the "dark matter" of certain parameter space), not necessarily the whole model and all its parameters.

3.2 Over-fitting Tendency with Feature Functions

Next, we generalize the fragility measure $\varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ to allow for transformations of parameters θ.

Definition 4 (Over-fitting Tendency Measure with Feature Functions)**.** *Let f be a $\mathbb{R}^{D_\Theta} \to \mathbb{R}^{D_f}$ continuously differentiable mapping with $1 \leq D_f \leq D_\Theta$. Then, we define*

$$\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) \equiv \int d_{S_\mathbb{Q}}\{f(\theta); \mathbf{x^n}, \mathbf{y^n}\}\pi_\mathbb{P}(\theta|\mathbf{x^n})d\theta, \tag{22}$$

$$\text{where} \quad d_{S_\mathbb{Q}}\{f(\theta); \mathbf{x^n}, \mathbf{y^n}\} = \inf_{(\tilde{\theta},\tilde{\psi}):f(\tilde{\theta})=f(\theta)} \widehat{J}_{n,S_\mathbb{Q}}(\tilde{\theta}, \tilde{\psi}) - \widehat{J}_{n,S_\mathbb{Q}}(\hat{\theta}^\mathbb{Q}, \hat{\psi}^\mathbb{Q}). \tag{23}$$

Here $\widehat{J}_{n,S_\mathbb{Q}}(\theta, \psi)$ is the J-distance defined in (4), and $(\hat{\theta}^\mathbb{Q}, \hat{\psi}^\mathbb{Q})$ is the GMM estimator.

Transforming the original parameter vector is useful, for example, if one wants to measure model's robustness with respect to a low-dimensional subset in the parameter space. For instance, to measure model robustness with respect to the first D_f elements of θ ($D_f < D_\Theta$), we set $f(\theta) = \mathbf{F}\theta$, where $\mathbf{F} = \nabla f(\theta) \equiv \left[I_{D_f}, O_{D_f \times (D_\Theta - D_f)}\right]$. In the special case of $f(\theta) = \mathbf{F}\theta$, with $\mathbf{F}$ being an arbitrary full-rank $D_\Theta \times D_\Theta$ matrix, we recover the original overall fragility measure, $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) = \varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$.[10]

3.3 Justification of Fisher Fragility Measures

The econometric over-fitting tendency measures in Sections 3.1 and 3.2 are built on the basis of well-established statistical over-fitting tendency measures. In this subsection, we show that our Fisher fragility measure fundamentally captures a particular structural component specification's over-fitting risk. In practice, computation of the over-fitting tendency measures $\varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ and $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ may be complicated by the complex form of the likelihood function of the baseline model, the curse of dimensionality induced by high-dimensional parameter spaces, and the additional minimization problem involved in the definition of the generalized measure. In contrast, the computation of our Fisher fragility measure is simple and an eigen-decomposition is derived.

The over-fitting tendency measure is connected to the original Fisher fragility measure defined in Section 2.2, as we show in Theorem 1. To derive the theoretical results, we require certain regularity conditions to discipline the behavior of the data. Specifically, the regularity conditions we choose are influenced by three major considerations. First, our assumptions are chosen to allow processes of sequential dependence, which should be relevant to inter-temporal asset pricing models. Second, our assumptions are required to meet the analytical tractability. Third, our assumptions are sufficient conditions in the sense that we are not trying to provide the weakest conditions to guarantee the theoretical results work.

We now introduce the regularity conditions for establishing the theoretical results in this paper. These regularity conditions are all standard in the econometric literature

[10]A similar monotonicity property to Proposition 19 applies to $\varrho_o^f(\theta_0, \psi_0; \mathbf{x^n}, \mathbf{y^n})$. Let $\widetilde{f} = [f, f_1]'$, where f and f_1 are continuously differentiable and $D_{\widetilde{f}} \le D_\Theta$. Then $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) \le \varrho_o^{\widetilde{f}}(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$.

of GMM. More detailed discussions and remarks can be found in Appendix A. The process $\{\mathbf{x_t}, \mathbf{y_t}\}$ is assumed to be a strictly stationary Markov process (Assumption A1). Specifically, we assume that the baseline moments to be included in the full vector of moment functions $g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t})$ for fragility assessment. The moment functions in $g_{\mathbb{Q}}$ are twice continuously differentiable, and the partial derivatives with respect to parameters satisfy standard dominance conditions. The times series satisfy the uniform mixing condition as in Newey (1985a) and White and Domowitz (1984). The mixing condition (Assumption A2) and the dominance condition (Assumption A3) are needed for the ULLN as in White and Domowitz (1984). They also imply the moment continuity of stochastic functions $g_{\mathbb{Q}}$, as well as their derivatives, adopted by Hansen (1982). Lastly, we need identification conditions to guarantee that the minimization problem in (5) has a unique solution asymptotically (Assumptions A4 and A5). A standard sufficient condition for GMM identification is that the covariance matrix $S_{\mathbb{Q}}$ is positive definite, the moment conditions (1) and (3) are satisfied only at θ_0, and the Fisher information matrix $\mathbf{I}_{\mathbb{Q}}(\theta_0, \psi_0)$ is non-singular. The GMM identification condition is standard in GMM literature.[11] The prior $\pi(\theta)$ is twice continuously differentiable and positive (Assumption A6). The proof of the following theorem and its corollaries are in Appendix B.

Theorem 1. *Consider a feature function $f : \mathbb{R}^{D_\Theta} \to \mathbb{R}^{D_f}$ with $\mathbf{v} = \nabla f(\theta_0)$ being the $D_f \times D_\Theta$ Jacobian matrix. Suppose the regularity conditions above (Assumptions A1 - A7 in Appendix A) hold. Then $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ converges in distribution to*

$$\operatorname*{wlim}_{n\to\infty} \varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) = \varrho^{\mathbf{v}}(\theta_0|\psi_0) + \sum_{i=1}^{D_f} \left[\lambda_i(\mathbf{v}) - 1\right] \chi^2_{1,i}, \tag{24}$$

where $\chi^2_{1,i}$'s are i.i.d. chi-squared random variables with 1 degree of freedom, and $\lambda_i(\mathbf{v})$'s are eigenvalues of $\Pi_0(\mathbf{v})$ defined in (13). Here "wlim" is the operator denoting a limit variable for convergence in distribution (weak convergence).

[11] See, e.g., Hansen (1982, Assumptions 3.4 - 3.6) and Newey (1985a, Assumptions 3 and 7).

Theorem 1 shows that the asymptotic distribution of over-fitting tendency measures can be characterized by the Fisher fragility measure $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ and the distribution of eigenvalues $\lambda_i(\mathbf{v})$ given their sum $\varrho^{\mathbf{v}}(\theta_0|\psi_0) = \sum_{i=1}^{D_f} \lambda_i(\mathbf{v})$ is kept constant. This result immediately leads to the following two corollaries.

Corollary 1. *Suppose the regularity conditions of Theorem 1 hold, then $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ and $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ are asymptotically related:*

$$\mathbb{E}\left[\operatorname*{wlim}_{n\to\infty} \varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})\right] = 2\varrho^{\mathbf{v}}(\theta_0|\psi_0) - D_f, \tag{25}$$

where $\mathbb{E}_{\mathbb{Q}}$ stands for the expectation under the distribution of the full model $\mathbb{Q}$.

In principle, the relationship of (25) shows that the Fisher fragility measure $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ captures the average tendency of over-fitting. It says, with a large sample size, the econometrician can use re-sampling methods such as bootstrap based on the sample $(\mathbf{x^n}, \mathbf{y^n})$ to estimate the average measure of over-fitting tendency. The limiting result in (25) guarantees that $2\varrho^{\mathbf{v}}(\theta_0|\psi_0) - D_f$ provides a reasonable approximation for such average measure of over-fitting tendency when sample size is large. As a special case where the feature function is the identical mapping, it holds that $\mathbb{E}\left[\operatorname{wlim}_{n\to\infty} \varrho_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})\right] = 2\varrho(\theta_0|\psi_0) - D_\Theta$.

The Fisher fragility $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ is a model property and does not depend on the sample. This measure focuses on local departures from the true parameter vector θ_0. Corollary 1 shows that it captures the average model over-fitting tendency as sample size goes large.

Theorem 1 also suggests that the likelihood of getting an extremely poor out-of-sample performance (i.e. an extremely large $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$) is determined by the distribution of the eigenvalues $\lambda_i(\mathbf{v})$ which decompose $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$ into subspaces.

Corollary 2. *Suppose the regularity conditions of Theorem 1 hold and the largest eigenvalue is $\lambda_1(\mathbf{v})$. Then, the tail probability of the limiting variable converges to zero* *at the exponential rate related negatively to* $\lambda_1(\mathbf{v})$*:*

$$\lim_{x\to\infty}\frac{1}{x}\ln\mathbb{P}\left\{\operatorname*{wlim}_{n\to\infty}\varrho_o^f(\theta_0|\psi_0,\mathbf{x^n},\mathbf{y^n})>x\right\}=-\frac{1}{2\left[\lambda_1(\mathbf{v})-1\right]} \tag{26}$$

Corollary 2 shows that the tail probability of the limiting variable converges to zero faster when the largest eigenvalue $\lambda_1(\mathbf{v})$ is smaller. For a given level of $\varrho^{\mathbf{v}}(\theta_0|\psi_0)$, which captures the *average* tendency of overfitting, a heavily skewed distribution of eigenvalues $\lambda_i(\mathbf{v})$ results in a larger value of $\lambda_1(\mathbf{v})$. Then, the tail of the distribution of $\varrho_o^f(\theta_0|\psi_0,\mathbf{x^n},\mathbf{y^n})$ is heavier and the probability of overfitting the data to an extremely large degree, is higher.

4 Applications

In this section, we illustrate and implement the Fisher fragility measure in the context of two widely studied asset pricing models. The first example is a rare disaster model, for which we compute the fragility measure analytically. The second example is a long-run risk model. We use this example to demonstrate how one can diagnose the sources of fragility in a more complex model and deal with nuisance parameters in measuring model fragility (or "dark matter").

4.1 Disaster Risk Model

Rare economic disasters are a natural source of "dark matter" in asset pricing models. It is difficult to evaluate the likelihood and the magnitude of rare disasters statistically. Yet, agents' aversion to large disasters can have an economically large effect on asset prices.[12]

We consider a disaster risk model similar to Barro (2006). The structural model describes the log growth rate of aggregate consumption g_t and the excess log return

[12] See the early work by Rietz (1988), and recent developments by Barro (2006), Gabaix (2012), Martin (2012), Wachter (2013), and Collin-Dufresne, Johannes, and Lochstoer (2016), among others.

on the market portfolio r_t. There are two regimes characterized by state variable z_t: the disaster regime ($z_t = 1$) and the outside-of-disaster regime ($z_t = 0$). The variable z_t has an i.i.d. Bernoulli distribution taking $z_t = 1$ with probability p, independent of the other variables. The realizations of z_t are observable. Outside of disasters ($z_t = 0$), g_t follows a normal distribution $N(\mu, \sigma^2)$; in a disaster state ($z_t = 1$), $g_t = -v_t$ where the log of decline in consumption v_t follows a truncated exponential distribution with density $\mathbf{1}\{v_t > \underline{v}\}\lambda e^{-\lambda(v_t - \underline{v})}$ with the lower bound for disaster size equal to $\underline{v}$ and the average disaster size equal to $\underline{v} + 1/\lambda$.

Now, we first impose the assumption on how excess log returns on market portfolio r_t depend on consumption growth g_t. More precisely, the joint distribution of (g_t, r_t) is time-varying contingent on the underlying disaster state z_t. When the economy is outside of disasters ($z_t = 0$), g_t and r_t are jointly normal:

$$r_t = \eta + \rho\frac{\tau}{\sigma}(g_t - \mu) + \sqrt{1-\rho^2}\tau u_t, \tag{27}$$

where u_t are independent standard normal shocks, and thus $\rho\frac{\tau}{\sigma}$ is the leverage factor in non-disaster states. When the economy is in a disaster state ($z_t = 1$), the excess log return is linked to the sizable decline in consumption with a leverage factor $b > 0$. In addition, we add an independent standard normal shock ε_t to r_t so that r_t and g_t are imperfectly correlated in a disaster state:

$$r_t = bg_t + \nu\varepsilon_t, \quad \text{with } \nu > 0. \tag{28}$$

Second, we impose the optimization restriction to the parameters of joint dynamics. The representative agent has a separable, constant relative risk aversion utility function $\sum_{t=0}^{\infty} \delta_{\mathrm{D}}^t c_t^{1-\gamma_{\mathrm{D}}}/(1-\gamma_{\mathrm{D}})$, where $\gamma_{\mathrm{D}} > 0$ is the coefficient of relative risk aversion. The log equity premium is $\mathbb{E}[r_t] = (1-p)\eta - pb\left(\underline{v} + 1/\lambda\right)$ where the Euler equation implies that η is a function of the other parameters (see Appendix D for details):

$$\eta = \gamma_{\text{D}}\rho\sigma\tau - \frac{\tau^2}{2} + \ln\left[1 + e^{\gamma_{\text{D}}\mu - \frac{\gamma_{\text{D}}^2\sigma^2}{2}}\Delta(\lambda)\frac{p}{1-p}\right] \tag{29}$$

$$\approx \gamma_{\text{D}}\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma_{\text{D}}\mu - \frac{\gamma_{\text{D}}^2\sigma^2}{2}}\Delta(\lambda)\frac{p}{1-p}, \tag{30}$$

where

$$\Delta(\lambda) = \lambda\left(\frac{e^{\gamma_{\text{D}}\underline{v}}}{\lambda - \gamma_{\text{D}}} - \frac{e^{\frac{\nu^2}{2} + (\gamma_{\text{D}} - b)\underline{v}}}{\lambda + b - \gamma_{\text{D}}}\right). \tag{31}$$

Equation (29) provides a cross-equation restriction among the consumption growth g_t, the disaster state z_t, and the excess log return of the market portfolio r_t. The first two terms on the right hand side give the market risk premium due to Gaussian consumption shocks. The third term is due to the disaster risk premium. We need $\lambda > \gamma_{\text{D}}$ for the risk premium to be finite, which sets an upper bound for the average disaster size and dictates how heavy the tail of the disaster size distribution can be.

The fact that the log equity premium $\mathbb{E}[r_t]$ explodes as λ approaches the value of γ_{D} is a crucial feature for our analysis. Even when we consider extremely rare disasters (very small p), we can still generate an arbitrarily large risk premium $\mathbb{E}[r_t]$ by making the average disaster size sufficiently large (lowering λ towards γ_{D}). Extremely rare and large disasters are difficult to rule out based on standard statistical tests. Below we illustrate how our fragility measure can detect fragility in models with such features.

Fisher fragility measure

Equations (27) – (31) together specify the full structural model. We set the baseline model to be the statistical model for rare disasters $\mathbf{x_t} = (z_t, v_t)$. Thus, the baseline parameters are $\theta = (p, \lambda)$. To focus our discussion on the rare disasters, we treat the other parameters $\phi = (\gamma_{\text{D}}, \mu, \sigma, \underline{v}, \tau, \rho, b, \nu)$ as auxiliary parameters fixed at known values. They are thus part of the functional-form specification in economic theories whose fragility is gauged. Naturally, $\mathbf{y_t} = (g_t, r_t)$. This treatment is reasonable since the key structural components of asset pricing models include not only preferences but also joint distribution of fundamentals and returns (see, e.g., Hansen and Singleton, 1983). And importantly, this simplifying treatment allows us to obtain a simple closed-form expression for the Fisher fragility measure. The nuisance parameter vector ψ is empty here.

Based on the approximation (30), the Fisher fragility measure is approximately (see Appendix D for details):

$$\varrho(p,\lambda) \approx 2 + \frac{p\Delta(\lambda)^2 + p(1-p)\lambda^2\dot{\Delta}(\lambda)^2}{(1-\rho^2)\tau^2(1-p)^2} e^{2\gamma_{\mathrm{D}}\mu - \gamma_{\mathrm{D}}^2\sigma^2}, \tag{32}$$

where $\dot{\Delta}(\lambda)$ is the first derivative of $\Delta(\lambda)$,

$$\dot{\Delta}(\lambda) = -\frac{e^{\gamma_{\mathrm{D}}\underline{v}}\gamma_{\mathrm{D}}}{(\lambda-\gamma_{\mathrm{D}})^2} + \frac{e^{(\gamma_{\mathrm{D}}-b)\underline{v}}(\gamma_{\mathrm{D}}-b)}{(\lambda-\gamma_{\mathrm{D}}+b)^2} e^{\nu^2/2}. \tag{33}$$

The one-dimensional worst-case asymptotic fragility measure is $\varrho^1(p,\lambda) = \varrho(p,\lambda) - 1$.

As Equation (29) shows, $\Delta(\lambda)$ and $\dot{\Delta}(\lambda)$ are related to the sensitivity of η to the disaster probability p and disaster size parameter λ, respectively. When λ approaches the value of γ_{D}, both $\Delta(\lambda)$ and $\dot{\Delta}(\lambda)$ approach infinity. Thus, disaster risk models with high average disaster size are fragile according to our measure.

Quantitative analysis

In our quantitative analysis, we use annual real per-capita consumption growth (nondurables and services) from the NIPA and returns on the CRSP value-weighted market portfolio for the period of 1929 to 2011. We fix the auxiliary parameters μ, σ, ν, τ and ρ at the values of the corresponding moments of the empirical distribution of consumption growth and excess stock returns: $\mu = 1.87\%$, $\sigma = 1.95\%$, $\tau = 19.14\%$, $\nu = 34.89\%$ and $\rho = 59.36\%$. The lower bound for disaster size is $\underline{v} = 7\%$. The leverage parameter b is 3. In Figure 2, we plot the 95% and 99% confidence regions for (p,λ) based on the baseline model.

The 95% confidence region for (p,λ) is quite wide. For low values of the disaster

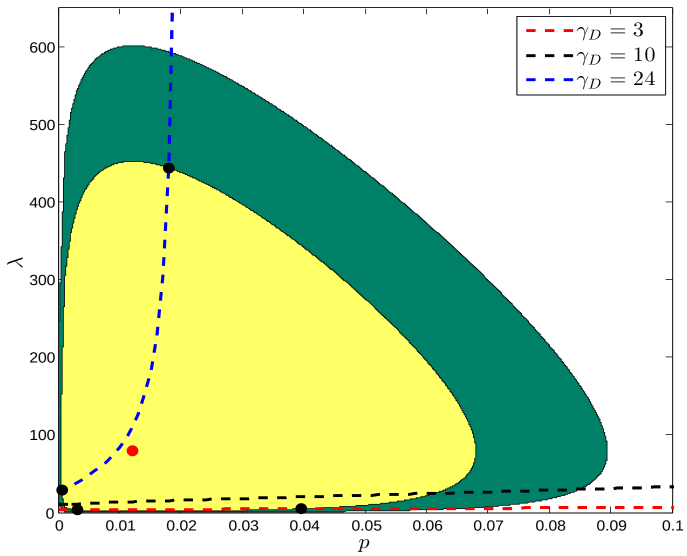

Figure 2: The 95% and 99% confidence regions of (p, λ) for the unconstrained model and the equity premium isoquants implied by the asset pricing constraint (29) for $\gamma_D = 3, 10, 24$. The maximum likelihood estimates are $(\widehat{p}_{\mathrm{ML}}, \widehat{\lambda}_{\mathrm{ML}}) = (0.0122, 78.7922)$.

probability p, the baseline model has little power to reject models with a wide range of average disaster size values (λ). Figure 2 also shows the equity premium isoquants for different levels of relative risk aversion: lines with the combinations of p and λ required to match the unconditional equity premium of 5.09% for a given value of γ_D. The fact that these isoquants all intersect with the 95% confidence region implies that even for low risk aversion ($\gamma_D = 3$), there exist many combinations of (p, λ) that not only match the observed equity premium, but also are "consistent with the macro data" in a sense that they cannot be rejected by the macro data based on standard statistical tests. In the remainder of this section, we refer to a calibration of (p, λ) that is within the 95% confidence region as an "acceptable calibration."[13]

While it is difficult to distinguish among a wide range of calibrations using standard statistical tools based on the macro data, these calibrated models differ significantly

[13] 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 historical sample, hence requiring the model to generate high equity premium using moderate disasters.

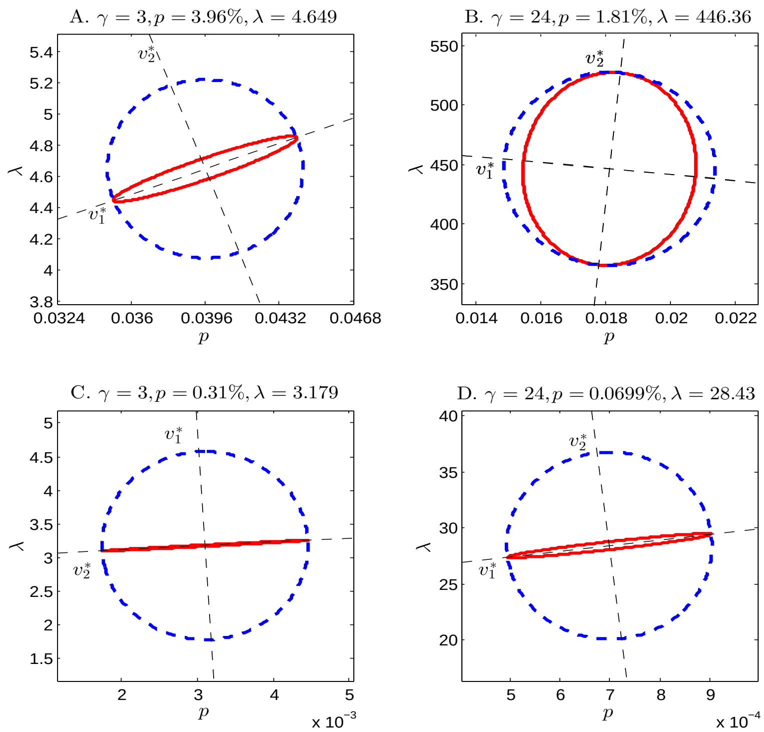

Figure 3: 95% confidence regions for the asymptotic distribution of the MLEs for four "acceptable calibrations." In Panels A through D, the Fisher fragility measures are $\varrho(p, \lambda) = 75.03$, 2.49, 1.78×10^4, and 5.61×10^2 respectively.

based on our fragility measures. We focus on four alternative calibrations, as denoted by the four points located at the intersections of the equity premium isoquants ($\gamma_{\mathrm{D}} = 3$ and 24) and the boundary of the 95% confidence region in Figure 2. For $\gamma_{\mathrm{D}} = 3$, the two points are ($p = 3.96\%, \lambda = 4.649$) and ($p = 0.31\%, \lambda = 3.179$). For $\gamma_{\mathrm{D}} = 24$, the two points are ($p = 1.81\%, \lambda = 446.36$) and ($p = 0.0699\%, \lambda = 28.43$).

With only two parameters in θ, we can illustrate the worst-case asymptotic fragility measure by plotting the asymptotic confidence regions for (p, λ) in the baseline model and the structural model, as determined by the respective information matrices $\mathbf{I}_{\mathbb{P}}(\theta)$ and $\mathbf{I}_{\mathbb{Q}}(\theta)$.[14] In each panel of Figure 3, the largest dash-line circle is the 95% confidence region for (p, λ) under the baseline model. The smaller solid-line ellipse is the 95%

[14] In fact, we use all the score functions of likelihoods to construct the moments, so the optimal GMM estimation is asymptotically equivalent to the MLE in our analysis of this disaster risk model.

confidence region for (p, λ) under the structural model. The reason that the confidence region under the structural model is smaller than that under the baseline model is that the GMM moments in the structural model contain both the moments for the baseline model and the moments (cross-equation restrictions) imposed by the structural component under fragility assessment. In this example, the two confidence regions coincide[15] in the direction of v_2^* and differ the most in the direction of v_1^*. Moreover, with enough extra data, the confidence region for the unconstrained estimator can be made small enough to reside within the confidence region of the constrained estimator.

In Panel A of Figure 3, with $\gamma_{\mathrm{D}} = 3, p = 3.96\%, \lambda = 4.649$, $\varrho(p, \lambda) = 75.07$ and $\varrho^1(p, \lambda) = 74.07$. This means that under the baseline model, we need to increase the amount of consumption data by a factor of 74.07 to match or exceed the precision in estimation of any linear combination of p and λ afforded by the equity premium constraint. Panels C and D of Figure 3 correspond to the calibrations with "extra rare and large disasters." For $\gamma_{\mathrm{D}} = 3$ and 24, $\varrho^1(p, \lambda)$ rises to 1.78×10^4 and 5.60×10^2, respectively. If, in Panel B of Figure 3, we raise γ_{D} to 24 while changing the annual disaster probability to 1.81% and lowering the average disaster size to 7.002% ($\lambda = 446.36$), $\varrho^1(p, \lambda)$ drops to 1.49. The reason is that by raising the risk aversion coefficient we are able to reduce the average disaster size.

So far, we have been examining the fragility of a specific calibrated structural model. We can also assess the fragility of a general class of models, relative to the baseline model of rare disasters, by plotting the distribution of $\varrho(\theta)$ based on a particular distribution of θ. For example, if econometricians are interested in fragility of a class of disaster risk models where the auxiliary parameters ϕ are fixed at given levels ϕ_0 and the uncertainty of baseline parameters θ is explicitly taken into account, the posterior for (p, λ) under the structural model (i.e., constrained posterior distribution) denoted by $\pi_{\mathbb{Q}}(\theta|\mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}})$ is proposed to be used as the distribution of θ.[16] Since the

[15] This is not true in general. When localized, the deterministic cross-equation restriction from the equity premium in this model is a linear constraint. Thus, the parameter estimates are not affected along the direction of the constraint.

[16] The simulated sample of the constrained posterior $\pi_{\mathbb{Q}}(\theta|\mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}})$ used for generate Figure 4 are

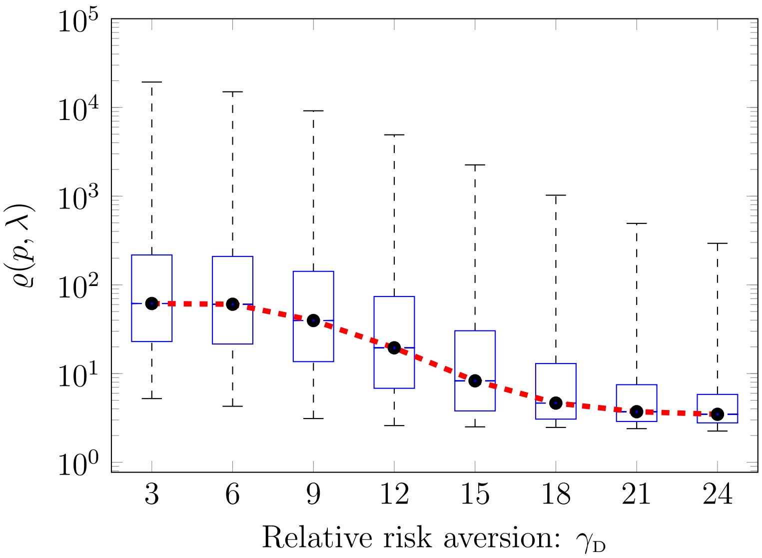

Figure 4: Distribution of the Fisher fragility measure $\varrho(p,\lambda)$ for different levels of risk aversion. For each γ_{D}, the boxplot shows the $1, 25, 50, 75$, and 99-th percentile of the distribution of $\varrho(p,\lambda)$ based on the constrained posterior for (p,λ).

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

We implement this idea in Figure 4. For each value of γ_{D}, the boxplot shows the $1, 25, 50, 75$, and 99-th percentile of the distribution of $\varrho(\theta)$ based on $\pi_{\mathbb{Q}}(\theta|\mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}})$. The Fisher fragility measures are higher when the levels of risk aversion are low. For example, for $\gamma_{\text{D}} = 3$, the $25, 50$, and 75-th percentile of the distribution of $\varrho(p,\lambda)$ are 23.0, 61.6, and 217.4, respectively. This is because a small value of γ_{D} forces the constrained posterior for θ to place more weight on "extra rare and large" disasters, which imposes particularly strong restrictions on the parameters (p,λ). As γ_{D} rises, the mass of the constrained posterior shifts towards smaller disasters, which imply lower information ratios. For $\gamma_{\text{D}} = 24$, the 25, 50, and 75-th percentile of the distribution of $\varrho(p,\lambda)$ drop to 2.8, 3.5, and 5.8, respectively.

simulated based on the Approximate Bayesian Computation (ABC) method, which is described in the online appendix.

4.2 Long-run risk model

In the second example, we consider a long-run risk model similar to Bansal and Yaron (2004) and Bansal, Kiku, and Yaron (2012). In the model, the representative agent has recursive preferences as in Epstein and Zin (1989) and Weil (1989) and maximizes his lifetime utility,

$$V_t = \left[(1-\delta_{\text{L}}) C_t^{1-1/\psi_{\text{L}}} + \delta_{\text{L}} \left(\mathbb{E}_t \left[V_{t+1}^{1-\gamma_{\text{L}}} \right] \right)^{\frac{1-1/\psi_{\text{L}}}{1-\gamma_{\text{L}}}} \right]^{\frac{1}{1-1/\psi_{\text{L}}}}, \tag{34}$$

where C_t is consumption at time t, δ_{L} is the rate of time preference, γ_{L} is the coefficient of risk aversion for timeless gambles, and ψ_{L} is the elasticity of intertemporal substitution when there is perfect certainty.

The log growth rate of consumption Δc_t, the conditional mean of consumption growth x_t, and the conditional volatility of consumption growth σ_t follow the process

$$\Delta c_{t+1} = \mu_c + x_t + \sigma_t \epsilon_{c,t+1} \tag{35a}$$

$$x_{t+1} = \rho x_t + \varphi_x \sigma_t \epsilon_{x,t+1} \tag{35b}$$

$$\widetilde{\sigma}_{t+1}^2 = \overline{\sigma}^2 + \nu(\widetilde{\sigma}_t^2 - \overline{\sigma}^2) + \sigma_w \epsilon_{\sigma,t+1} \tag{35c}$$

$$\sigma_{t+1}^2 = \max\left(\underline{\sigma}^2, \widetilde{\sigma}_{t+1}^2\right) \tag{35d}$$

where the shocks $\epsilon_{c,t}, \epsilon_{x,t}$, and $\epsilon_{\sigma,t}$ are *i.i.d.* $N(0,1)$ and mutually independent. The volatility process (35c) potentially allows for negative values of $\widetilde{\sigma}_t^2$. Following the literature, we impose a small positive lower bound $\underline{\sigma}$ (= 1 bps) for volatility σ_t in solutions and simulations. This negative volatility could be avoided in other ways. For example, the process of σ_t^2 can be modeled as a discrete-time version of the square root process. Next, the log dividend growth Δd_t follows the processes

$$\Delta d_{t+1} = \mu_d + \phi_d x_t + \varphi_{d,c} \sigma_t \epsilon_{c,t+1} + \varphi_{d,d} \sigma_t \epsilon_{d,t+1}, \tag{36}$$

where the shocks $\epsilon_{d,t}$ are *i.i.d.* $N(0,1)$ and mutually independent with the other shocks in (35a)–(35c).

From the consumption Euler equation, one can derive a linear approximation of the stochastic discount factor,

$$m_{t+1} = \Gamma_0 + \Gamma_1 x_t + \Gamma_2 \sigma_t^2 - \lambda_c \sigma_t \epsilon_{c,t+1} - \lambda_x \varphi_x \sigma_t \epsilon_{x,t+1} - \lambda_\sigma \sigma_w \epsilon_{\sigma,t+1}. \tag{37}$$

The formulae for the coefficients Γ_0, Γ_1, Γ_2, λ_c, λ_x, and λ_σ are standard in the long-run risk model literature and given in the online appendix. Moreover, the equilibrium excess (log) return follows

$$r^e_{m,t+1} = \mu^e_{r,t} + \beta_c \sigma_t \epsilon_{c,t+1} + \beta_x \sigma_t \epsilon_{x,t+1} + \beta_\sigma \sigma_w \epsilon_{\sigma,t+1} + \varphi_{d,d} \sigma_t \epsilon_{d,t+1}, \tag{38}$$

where the conditional average (log) excess return is

$$\mu^e_{r,t} = \lambda_c \beta_c \sigma_t^2 + \lambda_x \beta_x \varphi_x \sigma_t^2 + \lambda_\sigma \beta_\sigma \sigma_w^2 - \frac{1}{2} \sigma^2_{r_m,t}, \tag{39}$$

$$\text{where } \sigma^2_{r_m,t} = \beta_c^2 \sigma_t^2 + \beta_x^2 \sigma_t^2 + \beta_\sigma^2 \sigma_w^2 + \varphi_{d,d}^2 \sigma_t^2. \tag{40}$$

The expressions for β_c, β_x, and β_σ are also given in the online appendix.

There are stochastic singularities in the model. One example is that the excess log market return $r^e_{m,t+1}$ is a deterministic function of $\Delta c_{t+1}, \Delta d_{t+1}, x_{t+1}, x_t$, σ^2_{t+1}, and σ^2_t. Another example is that the market log price-dividend ratio $z_{m,t}$ is a deterministic function of x_t and σ_t^2. The moment-based methods such as GMM can focus on the marginal distribution and economic relationships targeted by the structural model. We focus on the marginal distribution of $(\Delta c_{t+1}, x_t, \sigma_t^2, \Delta d_{t+1}, r^e_{m,t+1})$, denoted by $\mathbb{Q}$ and specified by (35a) – (35d), (36), and (38) – (40), except that excess (log) returns in (38) is augmented by shocks $\varphi_r \sigma_t \epsilon_{r,t+1}$ with $\epsilon_{r,t}$ being i.i.d. standard normal variables and mutually independent with other variables. This is a standard approach in DSGE literature for dealing with stochastic singularity. In other words, although the structural model $\mathcal{Q}$ does not capture some part of the whole distribution $\mathbb{Q}$

Table 1: Benchmark Calibration for the Long-Run Risk Model

Preferences	δ_{L}	γ_{L}	ψ_{L}			
	0.9989	10	1.5			
Consumption	μ_c	ρ	φ_x	$\overline{\sigma}$	ν	σ_w
	0.0015	0.975	0.038	0.0072	0.999	$2.8e-6$
Dividends	μ_d	ϕ_d	$\varphi_{d,c}$	$\varphi_{d,d}$		
	0.0015	2.5	2.6	5.96		
Returns	φ_r					
	3.0					

(misspecified up to $\varphi_r \sigma_t \epsilon_{r,t+1}$), the moment conditions are correctly specified under the whole distribution $\mathbb{Q}$.[17]

Quantitative Analysis

We choose the model of consumption (35a)–(35d) as the baseline model $\mathcal{P}$. We assume that the econometrician observes the process for consumption, the latent variables x_t and σ_t^2, and the process for asset returns. We make the latent variables observable to be consistent with the postulated process for asset returns, which is derived assuming that these variables are observable.

Accordingly, the baseline parameters are $\theta = (\mu_c, \rho, \varphi_x, \overline{\sigma}^2, \nu, \sigma_w, \mu_d, \phi_d, \varphi_{d,c}, \varphi_{d,d})$ with $D_\Theta = 10$ and $\mathbf{x_t} = (\Delta c_{t+1}, x_t, \sigma_t^2, \Delta d_{t+1})$. By measuring the fragility of the long-run risk model relative to this particular baseline, we can interpret the fragility measure as quantifying the additional information that asset pricing restrictions provide for the consumption dynamics (in particular, the long-run risk components) relative to information contained in consumption data. We explicitly account for the uncertainty of preference parameters γ_{L} and ψ_{L} by including them into the nuisance parameter vector ψ. Thus, $\psi = (\gamma_{\text{L}}, \psi_{\text{L}})$. The extra data investigated by the full structural model $\mathcal{Q}$ are $\mathbf{y_t} = r^e_{m,t+1}$. Other parameters are included in the auxiliary parameter vector $\phi = (\delta_{\text{L}}, \varphi_r)$ which are fixed at known values, and these values are

[17] The moment conditions can be found in the online appendix.

part of the imposed functional-form specification of the structural component that is under the fragility assessment. It's worth noting that the joint dynamics between consumption growth processes and dividend growth are included in the baseline model, which is different from the disaster risk example. Thus, the fragility measure here is only about preferences and optimization specifications. Alternatively, including Δd_{t+1} into $\mathbf{y_t}$, our measure quantifies different model components' fragility.

The benchmark calibration of the model follows Bansal, Kiku, and Yaron (2012) and is summarized in Table 1. As Bansal, Kiku, and Yaron (2012) (Table 2, p. 194) show, the simulated first and second moments match the set of key asset pricing moments in the data reasonably well. The same is true for the alternative calibration in Table 1 (see the online appendix).

Let's first focus on Panel I of Table 2. It reports the fragility measures for preference parameters considered as nuisance parameters. The row (BC) reports the fragility measures for the benchmark calibration. The Fisher fragility measure is $\varrho = 276.3$, indicating a high level of model fragility. The worst-case 1-dimensional Fisher fragility measure is also high, $\varrho^1 = 196.3$, which implies that the sample size needs to be 195.3 times longer in order for the baseline model estimator to match the precision of the estimator for the full structural model in all 1-D directions.

The large size of ϱ^1 implies that the model under the benchmark calibration is highly sensitive to perturbations in the parameters in a single direction, as identified by v_1^* (i.e. the worst direction). However, this does not mean that one can discover the full scope of the fragility issue by examining individual parameters one at a time. We demonstrate this point by computing the fragility measure for each individual parameter $\varrho^{\mathbf{v}}$, where $\mathbf{v}$ is the appropriate standard basis vector $\mathbf{e}_i$ whose i-th element is one and other elements are zeros. As Panel I of Table 2 shows, the fragility measures for all the individual parameters are relatively small. While the measure is somewhat larger in magnitude for $\overline{\sigma}^2$ (the long-run variance of consumption growth) and ν (the persistence of conditional variance of consumption growth), all of the univariate measures are much smaller than ϱ and ϱ^1. Had we focused only on the sensitivity of

Table 2: Fragility Measures for the Long-Run Risk Models

Model	ϱ	ϱ^1	$\varrho^{\mathbf{v}}$									
			μ_c	ρ	φ_x	$\overline{\sigma}^2$	ν	σ_w	μ_d	ϕ_d	$\varphi_{d,c}$	$\varphi_{d,d}$
			I. Nuisance parameter vector ψ: $(\gamma_{\mathrm{L}}, \psi_{\mathrm{L}})$									
(BC)	276.3	196.3	1.0	1.1	1.0	48.9	97.8	1.0	1.0	3.4	1.0	1.0
(AC)	34.0	21.1	1.0	1.1	1.0	1.0	3.4	1.0	1.4	4.2	1.0	1.0
			II. Nuisance parameter vector ψ: empty									
(BC)	$3.58 \cdot 10^5$	$3.57 \cdot 10^5$	1.0	2.1	1.1	115.6	117.5	1.3	1.1	7.1	1.0	1.0
(AC)	323.3	287.7	1.0	2.5	1.0	1.0	6.3	1.0	1.9	31.3	1.0	1.0

Note: The direction corresponding to the worst-case 1-D fragility measure ϱ^1 for the benchmark calibration (BC) is given by $v_1^* = [0.000, 0.000, -0.000, 0.020, -0.001, 0.999, -0.001, 0.000, -0.000, 0.000]$. The alternative calibration (AC) has $\nu = 0.98$ and $\gamma_{\mathrm{L}} = 27$ with other parameters unchanged. In panel I, the uncertainty of preference parameters $(\gamma_{\mathrm{L}}, \psi_{\mathrm{L}})$ are accounted for; whereas in panel II, they are fixed as auxiliary parameters ϕ with nuisance parameter vector ψ empty.

the model's implications to individual parameters in θ, we would have missed the very high fragility of the full model and the large dark matter hidden in θ.

In comparison, Panel II of Table 2 reports fragility measures for the preference parameters being fixed. In such case, the preference specification is also part of the structural restrictions under fragility assessment. Thus, the fragility measures in Panel II are higher. Especially, the overall fragility ϱ and the worst-case 1-D fragility ϱ^1 increase dramatically from 276.3 to 3.58×10^5 and from 196.3 to 3.57×10^5, respectively.[18]

[18] In our example, we assume all long-run components (x_t and σ_t) are observable. If the long-run components are not observed and act as latent state variables, they need to be filtered (see Schorfheide, Song, and Yaron, 2014; Collin-Dufresne, Johannes, and Lochstoer, 2016), the model is likely to be more fragile. This is because there is more degrees of freedom to improve the in-sample fit of the asset pricing moments, while the persistence parameters ρ and ν are even more weakly identified by the joint dynamics of consumption growth and dividend growth processes. However, the models with latent state variables that endogenize the latent long-run risk in consumption growth (e.g. Hansen and Sargent, 2010) may be helpful to alleviate the fragility since the dynamic parameters of long-run components there can be endogenously linked to other fundamental variables for better identification.

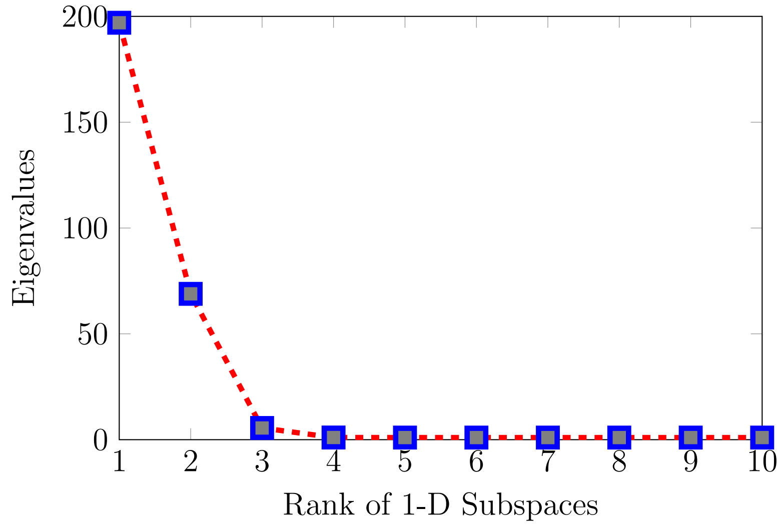

Figure 5: Eigenvalues for $\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}$ of the benchmark calibration.

Diagnosing the sources of fragility Besides measuring the fragility of the model, the Fisher fragility measures have provided a set of tools to diagnose the sources of fragility. First, the rankings of the eigenvalues of $\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{\frac{1}{2}}$ are informative. Each eigenvalue denotes the marginal contribution of a 1-D subspace to the overall fragility measure (see Definition 2 and Proposition 2). As Figure 5 shows, there are large differences between the eigenvalues. Model fragility along the worst direction in 1-dimensional subspaces, as captured by the leading eigenvalue, is 196.3, which accounts for over 71% of the total fragility. This result means that one can dramatically reduce the dimensionality (from 10 to 1) when analyzing the fragility of this model.

Second, the worst direction (i.e., the worst-case 1-dimensional subspace) is v_1^*. Knowing that the majority of the model fragility is concentrated in this direction, we can conveniently search for the fragile moments in the model by examining which moments are the most sensitive to the change in θ along the direction of v_1^*. For illustration, we focus on four moments from the long-run risk model, the risk loading and price of risk for volatility shocks $(\beta_\sigma\sigma_w, \lambda_\sigma\sigma_w)$, and for growth shocks $(\beta_x\sigma_t, \lambda_x\varphi_x\sigma_t)$. The conditional market excess return depends crucially on these

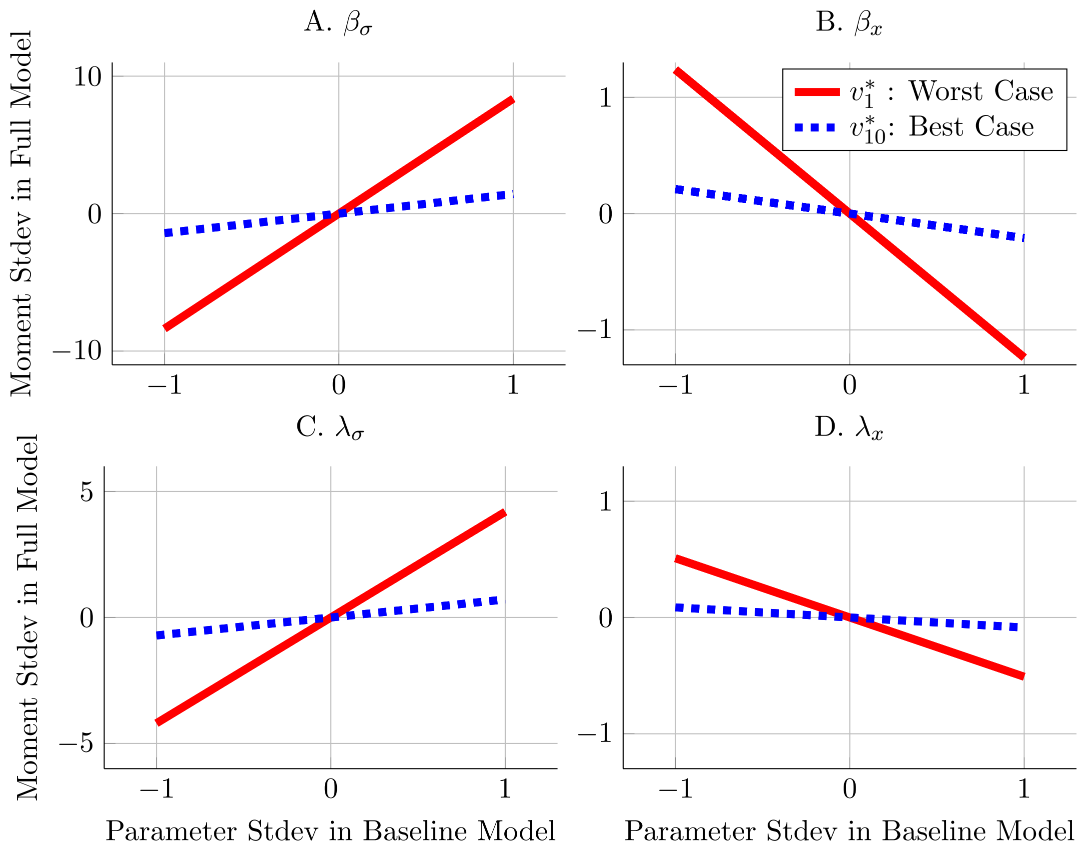

Figure 6: Sensitivity of return betas and risk prices with respect to the perturbation along the worst-case direction in the benchmark calibration with nuisance parameters.

moments (see Equation (39)).

In Figure 6, we plot the sensitivities of $\beta_\sigma, \lambda_\sigma, \beta_x$ and λ_x with respect to perturbations of θ along the worst direction v_1^* (solid line) and compare them to the sensitivities of the same set of moments to perturbations of θ along the best-case direction v_{10}^* (dash line). We measure the size of a perturbation of θ relative to the standard deviation of θ in the baseline model $\mathcal{P}$. We measure sensitivity of a moment as the change in the moment normalized by the moment's standard deviation in the structural model $\mathcal{Q}$.

The risk loading and the price of risk for volatility shocks are both highly sensitive to changes in θ along the direction of v_1^*, while the corresponding sensitivities to changes in θ along the direction of v_{10}^* are all very low in comparison. For example, a one standard deviation change in θ along the direction of v_1^* can lead to a 9-standard deviation change in β_σ under the full structural model $\mathcal{Q}$. Thus, an important source of fragility of the long-run risk model based on the benchmark calibration is in the risk exposure of the market portfolio to volatility shocks. If the true value of θ is slightly different from the benchmark calibration along the direction v_1^*, this version of the long-run risk model will perform poorly at explaining the relation between asset returns and volatility shocks out of sample.

Finally, we can further trace the sources of fragility by examining how λ_σ and β_σ are determined. For example, the fact that the persistence parameter for the conditional variance of consumption growth, ν, is close to 1, makes both β_σ and λ_σ sensitive to changes in θ. This motivates us to consider an alternative calibration (row (AC) of Table 2) with a smaller value for ν. Specifically, we change ν from 0.999 to 0.98, and simultaneously raise the coefficient of relative risk aversion γ from 10 to 27 in order to match the unconditional equity premium as in the benchmark calibration. The rest of the parameters are unchanged. This alternative calibration produces asset pricing moments largely similar to those in the benchmark calibration. However, based on our fragility measures, the alternative calibration is much less fragile compared to the benchmark calibration. As Panel I of Table 2 shows, under the alternative calibration (row (AC)), ϱ drops from 276.3 to 34, and ϱ^1 drops from 196.3 to 21.1.

5 Conclusion

In this paper, we propose a new measure of model fragility by quantifying a model's tendency of in-sample over-fitting. We formally connect the fragility of structural models to the informativeness of the cross-equation restrictions imposed on the parameters. We also provide a tractable asymptotic approximation to the fragility measure, which helps with diagnosing sources of model fragility.

Our methodology has a broad range of applications. In addition to the examples of applications in asset pricing that we consider in this paper, our measure can be used to assess robustness of structural models in many other areas of economics, such as structural industrial organization (IO) and structural corporate finance.

References

Akaike, H., 1973, “Information theory and an extension of the maximum likelihood principle,” in *Second International Symposium on Information Theory (Tsahkadsor, 1971)* . pp. 267–281, Akadémiai Kiadó, Budapest.

Ando, T., 2007, “Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models,” *Biometrika*, 94, 443–458.

Andrews, D. W. K., 1991, “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation,” *Econometrica*, 59, 817–58.

Andrews, D. W. K., and J. C. Monahan, 1992, “An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator,” *Econometrica*, 60, 953–66.

Bansal, R., D. Kiku, and A. Yaron, 2012, “An Empirical Evaluation of the Long-Run Risks Model for Asset Prices,” *Critical Finance Review*, 1, 183–221.

Bansal, R., and A. Yaron, 2004, “Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles,” *Journal of Finance*, 59, 1481–1509.

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

Box, G. E. P., 1976, “Science and statistics,” *J. Amer. Statist. Assoc.*, 71, 791–799.

Bradley, R. C., 2005, “Basic properties of strong mixing conditions. A survey and some open questions,” *Probab. Surv.*, 2, 107–144, Update of, and a supplement to, the 1986 original.

Breusch, T., H. Qian, P. Schmidt, and D. Wyhowski, 1999, “Redundancy of moment conditions,” *Journal of Econometrics*, 91, 89–111.

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

Chen, H., W. W. Dou, and L. Kogan, 2017, “Supplement to “Measuring the ‘Dark Matter’ in Asset Pricing Models”,” Discussion paper.

Cheng, X., and Z. Liao, 2015, “Select the valid and relevant moments: An information-based {LASSO} for {GMM} with many moments,” *Journal of Econometrics*, 186, 443 – 464.

Chernozhukov, V., and H. Hong, 2003, “An MCMC approach to classical estimation,” *J. Econometrics*, 115, 293–346.

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

Collin-Dufresne, P., M. Johannes, and L. A. Lochstoer, 2016, “Parameter Learning in General Equilibrium: The Asset Pricing Implications,” *American Economic Review*, 106, 664–98.

Efron, B., 2004, “The estimation of prediction error: covariance penalties and cross-validation,” *J. Amer. Statist. Assoc.*, 99, 619–642, With comments and a rejoinder by the author.

Epstein, L., and S. Zin, 1989, "Substitution, Risk Aversion, and the Temporal Behavior of Consumption Growth and Asset Returns I: A Theoretical Framework," *Econometrica*, 57, 937–969.

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.

Fisher, R. A., 1922, "On the mathematical foundations of theoretical statistics," *Phil. Trans. R. Soc. Lond. A*, 222, 309–68.

Foster, D. P., and E. I. George, 1994, "The risk inflation criterion for multiple regression," *Ann. Statist.*, 22, 1947–1975.

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

Gelman, A., J. Hwang, and A. Vehtari, 2013, "Understanding predictive information criteria for Bayesian models," *Statsitics and Computing*.

Gentzkow, M., and J. M. Shapiro, 2013, "Measuring the Sensitivity of Parameter Estimates to Sample Statistics," Chicago booth working papers.

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

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., J. Heaton, and A. Yaron, 1996, "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics*, 14, 262–80.

Hansen, L., and R. Jagannathan, 1997, "Assessing Specification Errors in Stochastic Discount Factor Models," *Journal of Finance*, 52, 557–90.

Hansen, L. P., 1982, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–54.

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

Hansen, L. P., 2014, "Nobel Lecture: Uncertainty Outside and Inside Economic Models," *Journal of Political Economy*, 122, 945 – 987.

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.

Hansen, L. P., and T. J. Sargent, 2010, "Fragile beliefs and the price of uncertainty," *Quantitative Economics*, 1, 129–162.

Hansen, L. P., and K. J. Singleton, 1982, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–86.

Hansen, L. P., and K. J. Singleton, 1983, "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, 91, 249–65.

Hausman, J., R. Lewis, K. Menzel, and W. Newey, 2011, "Properties of the CUE estimator and a modification with moments," *Journal of Econometrics*, 165, 45 – 57.

Herbst, E. P., and F. Schorfheide, 2015, *Bayesian Estimation of DSGE Models*, Princeton University Press.

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

Kim, J.-Y., 2002, "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, 107, 175–193.

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.

Krasker, W. S., 1980, "The "peso problem" in testing the efficiency of forward exchange markets," *Journal of Monetary Economics*, 6, 269 – 276.

Lewis, K. K., 2008, "Peso problem," in Steven N. Durlauf, and Lawrence E. Blume (ed.), *The New Palgrave Dictionary of Economics*, Palgrave Macmillan, Basingstoke.

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.

Nakagawa, K., 2005, "Tail probability of random variable and Laplace transform," *Appl. Anal.*, 84, 499–522.

Newey, W. K., 1985a, "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229–256.

Newey, W. K., 1985b, "Maximum likelihood specification testing and conditional moment tests," *Econometrica*, 53, 1047–1070.

Newey, W. K., and K. D. West, 1987a, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–08.

Newey, W. K., and K. D. West, 1987b, "Hypothesis testing with efficient method of moments estimation," *Internat. Econom. Rev.*, 28, 777–787.

Pinsker, M. S., 1964, *Information and Information Stability of Random Variables and Processes* . originally published in Russian in 1960, San Francisco: Holden-Day, San Francisco.

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

Rogoff, K. S., 1980, "Essays on Expectations and Exchange Rate Volatility," Discussion paper.

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

Schorfheide, F., D. Song, and A. Yaron, 2014, "Identifying Long-Run Risks: A Bayesian Mixed-Frequency Approach," Nber working papers, National Bureau of Economic Research, Inc.

Schwarz, G., 1978, "Estimating the dimension of a model," *Ann. Statist.*, 6, 461–464.

Shen, X., and J. Ye, 2002, "Adaptive model selection," *J. Amer. Statist. Assoc.*, 97, 210–221.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde, 2002, "Bayesian measures of model complexity and fit," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64, 583–639.

Tibshirani, R., and K. Knight, 1999, "The Covariance Inflation Criterion for Adaptive Model Selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 529–546.

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

Veronesi, P., 2004, "The Peso problem hypothesis and stock market returns," *Journal of Economic Dynamics and Control*, 28, 707–725.

Wachter, J. A., 2013, "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?," *Journal of Finance*, 68, 987–1035.

Weil, P., 1989, "The Equity Premium Puzzle and the Risk-Free Rate Puzzle," *Journal of Monetary Economics*, 24, 401–421.

White, H., and I. Domowitz, 1984, "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143–61.

Ye, J., 1998, "On measuring and correcting the effects of data mining and model selection," *J. Amer. Statist. Assoc.*, 93, 120–131.

Zin, S. E., 2002, "Are behavioral asset-pricing models structural?," *Journal of Monetary Economics*, 49, 215–228.

Appendix

A Regularity Conditions for Theoretical Results

The regularity conditions we choose to impose on the behavior of the data are influenced by three major considerations. First, our assumptions are chosen to allow processes of sequential dependence. In particular, the processes allowed should be relevant to intertemporal asset pricing models. Second, our assumptions are required to meet the analytical tractability. Third, our assumptions are sufficient conditions in the sense that we are not trying to provide the weakest conditions or high level conditions to guarantee the results; but instead, we chose those regularity conditions which are relatively straightforward to check in practice.

Assumption A1 (Stationarity Condition)
We assume that the underlying time series (x_t, y_t) with $t = 1, \cdots, n$ follow an m_S-order strictly stationary Markov process. Thus, the marginal conditional density for x_t can be specified as $\pi_{\mathbb{P}}(x_t|\theta, x_{t-1}, \cdots, x_{t-m_S})$. Define the stacked vectors $\mathbf{x_t} = (x_t, \cdots, x_{t-m_S})^T$ and $\mathbf{y_t} = (y_t, \cdots, y_{t-m_S})^T$, then the marginal conditional density under $\mathbb{P}$ can be rewritten as $\pi_{\mathbb{P}}(\mathbf{x_t}; \theta)$. The stacked vectors $(\mathbf{x_t}, \mathbf{y_t})$ follow a first-order Markov process.

Assumption A2 (Mixing Condition)
The stationarity condition and the m-dependence condition imply that there exists constant $\lambda_D \geq 2d_D/(d_D - 1)$, where d_D is the constant in Assumption 3 (dominance condition), such that $(\mathbf{x_t}, \mathbf{y_t})$ for $t = 1, 2, \cdots, n$ is uniform mixing and there exists a constant $\bar{\phi}$ such that the uniform mixing coefficients satisfy

$$\phi(m) \leq \bar{\phi} m^{-\lambda_D} \quad \text{for all possible probabilistic models,}$$

where $\phi(m)$ is the uniform mixing coefficient. Its definition is standard and can be found, for example, in White and Domowitz (1984) or Bradley (2005).

Remark. *Following the literature (see, e.g. White and Domowitz, 1984; Newey, 1985b; Newey and West, 1987a), we adopt the mixing conditions as a convenient way of describing economic and financial data which allows time dependence and heteroskedasticity. The mixing conditions basically restrict the memory of a process to be weak, while allowing heteroskedasticity, so that large sample properties of the process are preserved. In particular, we employ the uniform mixing which is discussed in some detail by White and Domowitz (1984) where definition and its relationship with other type of mixing conditions can be found in the survey by Bradley (2005).*

Assumption A3 (Dominance Condition)
The function $g_{\mathbb{Q}}(\theta, \psi; \mathbf{x}, \mathbf{y})$ is twice continuously differentiable in (θ, ψ) almost surely. There exist dominating measurable functions $a_1(\mathbf{x}, \mathbf{y})$ and $a_2(\mathbf{x}, \mathbf{y})$, and constant $d_D > 1$, such that almost everywhere

$$\begin{aligned}
&|g_{\mathbb{Q}}(\theta, \psi; \mathbf{x}, \mathbf{y})|^2 \leq a_1(\mathbf{x}, \mathbf{y}), \quad ||\nabla g_{\mathbb{Q}}(\theta, \psi; \mathbf{x}, \mathbf{y})||_{\mathbb{S}}^2 \leq a_1(\mathbf{x}, \mathbf{y}),\\
&||\nabla^2 g_{\mathbb{Q},(i)}(\theta, \psi; \mathbf{x}, \mathbf{y})||_{\mathbb{S}}^2 \leq a_1(\mathbf{x}, \mathbf{y}), \quad \text{for } i = 1, \cdots, D_g,\\
&|q(\mathbf{x}, \mathbf{y})| \leq a_2(\mathbf{x}, \mathbf{y}), \quad |q(\mathbf{x_1}, \mathbf{y_1}, \mathbf{x_t}, \mathbf{y_t})| \leq a_2(\mathbf{x_1}, \mathbf{y_1}) a_2(\mathbf{x_t}, \mathbf{y_t}), \quad \text{for } t \geq 2,\\
&\int [a_1(\mathbf{x}, \mathbf{y})]^{d_D} a_2(\mathbf{x}, \mathbf{y}) \mathrm{d}x \mathrm{d}y < +\infty, \quad \int a_2(\mathbf{x}, \mathbf{y}) \mathrm{d}x \mathrm{d}y < +\infty,
\end{aligned}$$

where $|| \cdot ||_{\mathbb{S}}$ is the spectral norm of matrices.

Remark. *The dominating function assumption is widely adopted in the literature of generalized method of moments (Newey, 1985a,b; Newey and West, 1987a). The dominating assumption, together with the uniform mixing assumption and stationarity assumption, imply the stochastic equicontinuity* *condition (iv) in Proposition 1 of Chernozhukov and Hong (2003). In the seminal GMM paper by Hansen (1982), the moment continuity condition can also be derived from the dominance conditions.*

Remark. *Consider the assumption of the theoretical results in Section 3 where $g_{\mathbb{P}}(\theta; \mathbf{x}, \mathbf{y})$ contains all the score functions of well specified baseline likelihood. For each pair of j and k, it holds that for some constant $\zeta > 0$ and large constant $C > 0$,*

$$\mathbb{E}_{\mathbb{P}} \sup_{\theta \in \Theta} \left| \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ln \pi_{\mathbb{P}}(\mathbf{w}; \theta) \right|^{2+\zeta} < C, \quad \textit{and} \quad \mathbb{E}_{\mathbb{P}} \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_j} \ln \pi_{\mathbb{P}}(\mathbf{w}; \theta) \right|^{2+\zeta} < C, \quad \textit{and} \tag{41}$$

The dominance condition, together with the uniform mixing assumption and stationarity assumption, implies the stochastic equicontinuity condition (i) in Proposition 3 of Chernozhukov and Hong (2003).

Assumption A4 (Nonsingular Condition)
The Fisher information matrices $\mathbf{I}_{\mathbb{P}}(\theta)$ and $\mathbf{I}_{\mathbb{Q}}(\theta, \psi)$ are positive definite for all θ, ψ.

Remark. *It implies that the covariance matrices $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$ are positive definite, and the expected moment function gradients $G_{\mathbb{P}}(\theta)$ and $G_{\mathbb{Q}}(\theta, \psi)$ have full rank for all θ and ψ.*

Assumption A5 (Identification Condition)
The true baseline parameter vector θ_0 is identified by the baseline moment conditions in the sense that $\mathbb{E}_{\mathbb{P}}[g_{\mathbb{P}}(\theta; \mathbf{x})] = 0$ only if $\theta = \theta_0$. And, the true parameters (θ_0, ψ_0) of the full model is identified by the moment conditions in the sense that $\mathbb{E}_{\mathbb{Q}}[g_{\mathbb{Q}}(\theta, \psi; \mathbf{x}, \mathbf{y})] = 0$ only if $\theta = \theta_0$ and $\psi = \psi_0$.

Remark. *Consider the assumption of well-specified likelihood in Section 3. The continuous differentiability of moment functions, together with the identification condition, imply that the parametric family of distributions $\mathbb{P}_\theta$, as well as the moment conditions, are sound: the convergence of a sequence of parameter values is equivalent to the weak convergence of the distributions:*

$$\theta \to \theta_0 \Leftrightarrow \mathbb{P}_\theta \to \mathbb{P}_{\theta_0} \Leftrightarrow \mathbb{E}_{\mathbb{P}}[\ln(d\mathbb{P}_\theta / d\mathbb{P})] \to \mathbb{E}_{\mathbb{P}}[\ln(d\mathbb{P}_{\theta_0} / d\mathbb{P})] = 0. \tag{42}$$

Let $\gamma = (\theta, \psi)$ and $\gamma_0 = (\theta_0, \psi_0)$. The convergence of a sequence of parameter values is equivalent to the convergence of the moment conditions:

$$\gamma \to \gamma_0 \Leftrightarrow \mathbb{E}_{\mathbb{Q}}[g_{\mathbb{Q}}(\gamma; \mathbf{x}, \mathbf{y})] \to \mathbb{E}_{\mathbb{Q}_0}[g_{\mathbb{Q}}(\gamma_0; \mathbf{x}, \mathbf{y})] = 0. \tag{43}$$

Assumption A6 (Regular Bayesian Condition)
Suppose the parameter set is $\Theta \times \Psi \subset \mathbb{R}^{D_\Theta + D_\Psi}$ with Θ and Ψ being compact. And, the prior is absolutely continuous with respect to the Lebesgue measure with Radon-Nykodim density $\pi(\theta, \psi)$, which is twice continuously differentiable and positive. Denote $\overline{\pi} \equiv \max_{\theta \in \Theta, \psi \in \Psi} \pi(\theta, \psi)$ and $\underline{\pi} \equiv \min_{\theta \in \Theta, \psi \in \Psi} \pi(\theta, \psi)$. The probability measure defined by the limited-information posterior density $\pi_{\mathbb{Q}}(\theta | \mathbf{x^n}, \mathbf{y^n})$ is dominated by the probability measure defined by the baseline limited-information posterior density $\pi_{\mathbb{P}}(\theta | \mathbf{x^n})$, for almost every $\mathbf{x^n}, \mathbf{y^n}$ under $\mathbb{Q}_0$.

Remark. *Compactness implies total boundness. In our diaster risk model, the parameter set for the prior 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.*

Remark. *The concept of dominating measure here is the one in measure theory. More precisely, this regularity condition requires that for any measurable set which has zero measure under $\pi_{\mathbb{Q}}(\theta | \mathbf{x^n}, \mathbf{y^n})$, it must also have zero measure under $\pi_{\mathbb{P}}(\theta | \mathbf{x^n})$. This assumption is just to guarantee that $\mathbf{D}_{KL}(\pi_{\mathbb{Q}}(\theta | \mathbf{x^n}, \mathbf{y^n}) || \pi_{\mathbb{P}}(\theta | \mathbf{x^n}))$ to be well defined.*

Assumption A7 (Regular Feature Function Condition)
The feature function $f: \Theta \to \mathbb{R}$ is twice continuously differentiable. We assume that there exist $D_\Theta - 1$

twice continuously differentiable functions $f_2, \cdots, f_{D_\Theta}$ on Θ such that $F = (f, f_2, \cdots, f_{D_\Theta}) : \Theta \to \mathbb{R}^{D_\Theta}$ is a one-to-one mapping (i.e. injection) and $F(\Theta)$ is a connected and compact D_Θ-dimensional subset of $\mathbb{R}^{D_\Theta}$.

Remark. *A simple sufficient condition for the regular feature function condition to hold is that f is a proper and twice continuously differentiable function on $\mathbb{R}^{D_\Theta}$ and $\frac{\partial f(\theta)}{\partial \theta_{(1)}} > 0$ at each $\theta \in \mathbb{R}^{D_\Theta}$. 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 $\theta \in \mathbb{R}^{D_\Theta}$ and F is proper and twice differentiable mapping $\mathbb{R}^{D_\Theta} \to \mathbb{R}^{D_\Theta}$. According to the Hadamard's Global Inverse Function Theorem (e.g. Krantz and Parks, 2013), we know that F is a one-to-one mapping and $F(\Theta)$ is a connected and compact D_Θ-dimensional subset of $\mathbb{R}^{D_\Theta}$.*

B Proof of Theorem 1 And Its Corollaries

B.1 Asymptotic Normality of Posteriors

Proposition 4. *Under Assumptions A1 - A6 in Appendix A, it holds that*

$$\mathbf{D}_{KL}(\pi_{\mathbb{P}}(\theta|\mathbf{x^n})||N(\hat{\theta}^{\mathbb{P}}_{ML}, n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1})) \to 0 \quad \text{in } \mathbb{P}_n.$$

Proof. We extend the proof of Theorem 2.1 in Clarke (1999) which is under the i.i.d. condition. However, we have to adjust two parts of their proof, to extend the result to the case that the observations are time series with uniform mixing. The first part is to show that $\sup_{\theta\in\Theta} |\widehat{H}_{\mathbb{P},n}(\theta)| = O_p(1)$ where

$$\widehat{H}_{\mathbb{P},n}(\theta) \equiv -\frac{1}{n}\sum_{t=1}^{n} \ln \pi_{\mathbb{P}}(\mathbf{x_t}; \theta). \tag{44}$$

When n is large enough, we obtain that

$$\sup_{\theta\in\Theta} |\widehat{H}_{\mathbb{P},n}(\theta)| \leq 1 + \frac{1}{n}\sum_{t=1}^{n} \sup_{\theta\in\Theta} |\ln \pi_{\mathbb{P}}(\mathbf{x_t}; \theta)|.$$

Based on the mixing condition and the dominance condition, it follows from Theorem 2.3 of White and Domowitz (1984) that

$$\frac{1}{n}\sum_{t=1}^{n} \sup_{\theta\in\Theta} |\ln \pi_{\mathbb{P}}(\mathbf{x_t}; \theta)| \to \mathbb{E}_{\mathbb{P}} \sup_{\theta\in\Theta} |\ln \pi_{\mathbb{P}}(\mathbf{x_t}; \theta)| \quad \text{a.s.}$$

which further implies that $\sup_{\theta\in\Theta} |\widehat{H}_{\mathbb{P},n}(\theta)| = O_p(1)$. The second part is to show that

$$\int u^T u \left|\pi_{\mathbb{P}}(\hat{\theta}^{\mathbb{P}}_{\mathrm{ML}} + u/\sqrt{n}|\mathbf{x^n}) - \varphi_{\mathbb{P}}(u)\right| \mathrm{d}u \to 0 \quad \text{in} \quad \mathbb{P}_{0,n} \tag{45}$$

where $\varphi_{\mathbb{P}}(u) = \sqrt{\det \mathbf{I}_{\mathbb{P}}(\theta_0)/(2\pi)^{D_\Theta}} \exp\left[-\frac{1}{2}u^T\mathbf{I}_{\mathbb{P}}(\theta_0)u\right]$. In Clarke (1999), it shows that when $\mathbf{x_1}, \cdots, \mathbf{x_n}$ are i.i.d., the limit result (45) is satisfied under the regularity conditions in Assumptions A3 - A6. To extend this limit result to allow weak dependence, we appeal to Theorem 1 and Proposition 3 of Chernozhukov and Hong (2003) whose conditions are implied by Assumptions A1 - A6 in Appendix A. □

B.2 Proof of Theorem 1

In this section, we prove the result of Theorem 1. One major additional technical challenge, compared with standard large-sample inferences of GMM (see Hansen, 1982; Newey, 1985a,b), is that we need to establish uniform convergence and bounds for constrained GMM over a set of mis-specified moment constraints. Thus we need special treatments in our proof in establishing the uniform convergence and bounds.

Because of Assumption A7 (the regular feature function condition), as well as the fact that the definition of our "dark matter" measure (Definition 4) and regularity assumptions A1 - A6 in Appendix A are invariant under invertible and second-order smooth transformations of parameters, we can assume that $f(\theta) = \theta_1 \equiv (\theta_{(1)}, \cdots, \theta_{(D_f)})^T$ and hence $\nabla f(\theta) \equiv \mathbf{v} = \left[I_{D_f}, O_{D_f \times (D_\Theta - D_f)}\right]$, without loss of generality. That is, $f(\theta) = \theta_1 = \mathbf{v}\theta$. The remaining baseline parameters are summarized in $\theta_2 \equiv (\theta_{(D_f+1)}, \cdots, \theta_{(D_\Theta)})^T$. Thus, the over-fitting tendency measure with feature functions can be written as

$$\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) \equiv \int d_{S_\mathbb{Q}}\{\mathbf{v}\theta; \mathbf{x^n}, \mathbf{y^n}\}\pi_\mathbb{P}(\theta|\mathbf{x^n})\mathrm{d}\theta, \quad \text{where} \tag{46}$$

$$d_{S_\mathbb{Q}}\{\mathbf{v}\theta; \mathbf{x^n}, \mathbf{y^n}\} = \inf_{(\tilde{\theta}, \tilde{\psi}) \,:\, \mathbf{v}\tilde{\theta} = \mathbf{v}\theta} \widehat{J}_{n,S_\mathbb{Q}}(\tilde{\theta}, \tilde{\psi}) - \widehat{J}_{n,S_\mathbb{Q}}(\hat{\theta}^\mathbb{Q}, \hat{\psi}^\mathbb{Q}). \tag{47}$$

For clear exposition, we divide our whole proof into the following steps.

(1) A local reparametrization. Because it is an asymptotic equivalence result in a parametric setting, we follow the convention of asymptotic statistics (see, e.g. van der Vaart, 1998) to consider the local reparametrization:

$$(\theta, \psi) = (\hat{\theta}^\mathbb{P}, \hat{\psi}^\mathbb{P}) + \frac{1}{\sqrt{n}}(u, h). \tag{48}$$

Thus, the over-fitting tendency measure defined in (46) and (47) can be rewritten as

$$\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) \equiv \int d_{S_\mathbb{Q}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}\pi_\mathbb{P}(\hat{\theta}^\mathbb{P} + u/\sqrt{n})\mathrm{d}u, \quad \text{where} \tag{49}$$

$$\begin{aligned} d_{S_\mathbb{Q}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\} = \inf_{(\tilde{u}, \tilde{h}) \,:\, \mathbf{v}\tilde{u} = \mathbf{v}u} & \widehat{J}_{n,S_\mathbb{Q}}\left(\hat{\theta}^\mathbb{P} + \tilde{u}/\sqrt{n},\ \hat{\psi}^\mathbb{P} + \tilde{h}/\sqrt{n}\right) \\ & - \widehat{J}_{n,S_\mathbb{Q}}\left(\hat{\theta}^\mathbb{P} + \hat{u}^\mathbb{Q}/\sqrt{n},\ \hat{\psi}^\mathbb{P} + \hat{h}^\mathbb{Q}/\sqrt{n}\right). \end{aligned} \tag{50}$$

The transformed variable $u_1 = \mathbf{v}u$ is defined by $\mathbf{v}u = \sqrt{n}\left(\mathbf{v}\theta - \mathbf{v}\hat{\theta}^\mathbb{P}\right)$, and other transformed variables are defined analogously. The GMM estimator of the transformed variables are $(\hat{u}^\mathbb{Q}, \hat{h}^\mathbb{Q})$ such that $(\hat{\theta}^\mathbb{Q}, \hat{\psi}^\mathbb{Q}) = \left(\hat{\theta}^\mathbb{P} + \hat{u}^\mathbb{Q}/\sqrt{n}, \hat{\psi}^\mathbb{P} + \hat{h}^\mathbb{Q}/\sqrt{n}\right)$.

(2) Uniform quadratic bounds for the over-fitting gap $d_{S_\mathbb{Q}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}$. According to Assumption 3 (the dominance condition), the distance $d_{S_\mathbb{Q}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}$ is bounded by

$$0 \le d_{S_\mathbb{Q}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\} \le 4n\underline{\lambda}_{S_\mathbb{Q}}^{-1} \left|\widehat{g}_{\mathbb{Q},n}(\theta, \psi)\right|^2 \le 4\underline{\lambda}_{S_\mathbb{Q}}^{-1} \sum_{t=1}^n a_1(\mathbf{x_t}, \mathbf{y_t})$$

where $\underline{\lambda}_{S_\mathbb{Q}}$ is the smallest eigenvalue of $S_\mathbb{Q}$. The uniform upper bound is crude because it does not take the advantage of $\widehat{g}_{\mathbb{Q},n}(\theta, \psi)$ being close to zero when (θ, ψ) is in the local neighborhood of (θ_0, ψ_0).

Now, we provide much sharper quadratic bounds. First, it holds that,

$$d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\} \leq 2\widehat{J}_{n,S_{\mathbb{Q}}}\left(\hat{\theta}^{\mathbb{P}}+\frac{u}{\sqrt{n}},\hat{\psi}^{\mathbb{Q}}\right)-2\widehat{J}_{n,S_{\mathbb{Q}}}\left(\hat{\theta}^{\mathbb{P}}+\frac{\hat{u}^{\mathbb{Q}}}{\sqrt{n}},\hat{\psi}^{\mathbb{Q}}\right). \tag{51}$$

The inequality above is point by point for all sample $(\mathbf{x^n},\mathbf{y^n})$ and all $u \in \mathbb{R}^{D_\Theta}$.

The second-order Taylor expansion of $\widehat{J}_{n,S_{\mathbb{Q}}}\left(\hat{\theta}^{\mathbb{P}}+u/\sqrt{n},\hat{\psi}^{\mathbb{Q}}\right)$ around $(\hat{\theta}^{\mathbb{Q}},\hat{\psi}^{\mathbb{Q}})$ implies the following relationship:

$$\begin{aligned}&\widehat{J}_{n,S_{\mathbb{Q}}}\left(\hat{\theta}^{\mathbb{P}}+u/\sqrt{n},\hat{\psi}^{\mathbb{Q}}\right)-\widehat{J}_{n,S_{\mathbb{Q}}}\left(\hat{\theta}^{\mathbb{P}}+\hat{u}^{\mathbb{Q}}/\sqrt{n},\hat{\psi}^{\mathbb{Q}}\right)\\&\qquad=(u-\hat{u}^{\mathbb{Q}})^T\left\{\Gamma_\Theta\left[n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta_u,\hat{\psi}^{\mathbb{Q}})\right]\Gamma_\Theta^T\right\}(u-\hat{u}^{\mathbb{Q}})\end{aligned} \tag{52}$$

where θ_u lies on the segment between $\hat{\theta}^{\mathbb{P}}+u/\sqrt{n}$ and $\hat{\theta}^{\mathbb{Q}}=\hat{\theta}^{\mathbb{P}}+\hat{u}^{\mathbb{Q}}/\sqrt{n}$. The Hessian matrix has the expression

$$\begin{aligned}n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta_u,\hat{\psi}^{\mathbb{Q}}) &= 2\left[\nabla\widehat{g}_{\mathbb{Q},n}(\theta_u,\hat{\psi}^{\mathbb{Q}})\right]^T S_{\mathbb{Q}}^{-1}\left[\nabla\widehat{g}_{\mathbb{Q},n}(\theta_u,\hat{\psi}^{\mathbb{Q}})\right]\\&\quad+2\sum_{i=1}^{D_g}\nabla^2\hat{g}_{\mathbb{Q},n,(i)}(\theta_u,\hat{\psi}^{\mathbb{Q}})\left[\mathbf{e}_i^T S_{\mathbb{Q}}^{-1}\hat{g}_{\mathbb{Q},n}(\theta_u,\hat{\psi}^{\mathbb{Q}})\right]\end{aligned} \tag{53}$$

where $\mathbf{e}_i$ is a column vector with its i-th element equal to one and others equal to zeros, and $\hat{g}_{\mathbb{Q},n,(i)}(\theta_u,\hat{\psi}^{\mathbb{Q}})$ is the i-th element of $\hat{g}_{\mathbb{Q},n}(\theta_u,\hat{\psi}^{\mathbb{Q}})$.

According to the expression above and Assumption 3 (the dominance condition), there exist a sequence of integral nonnegative variables $D_{\mathbb{Q},n}(\mathbf{x^n},\mathbf{y^n})$ such that $D_{\mathbb{Q},n}(\mathbf{x^n},\mathbf{y^n})$'s first moments are uniformly bounded over n and

$$\left|\left|\Gamma_\Theta\left[n^{-1}\nabla^2 J_{n,S_{\mathbb{Q}}}(\theta_u,\hat{\psi}^{\mathbb{Q}};\mathbf{x^n},\mathbf{y^n})\right]\Gamma_\Theta^T\right|\right|_{\mathcal{S}} \leq D_{\mathbb{Q},n}(\mathbf{x^n},\mathbf{y^n}), \quad \text{for all } u. \tag{54}$$

From (51), (52), and (54), the over-fitting gap $d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}$ has the upper bound

$$d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\} \leq D_{\mathbb{Q},n}(\mathbf{x^n},\mathbf{y^n})|u-\hat{u}^{\mathbb{Q}}|^2, \quad \text{for all } u. \tag{55}$$

(3) Uniform lower bounds for the Hessian matrix of J-distance $n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta,\psi)$. Now we show that there exist constants $\underline{\lambda}_J>0$ and $\delta>0$ such that, with probability converges to one, the smallest eigenvalue of Hessian matrix $n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta,\psi)$ is bigger than $\underline{\lambda}_J$ uniformly over $\mathcal{N}_0(\delta)\equiv\left\{(\theta,\psi):\ |\theta-\theta_0|^2+|\psi-\psi_0|^2\leq\delta^2\right\}$. In other words, there exist $\underline{\lambda}_J>0$ and $\delta>0$ such that

$$\mathbb{Q}_{0,n}\left\{\text{All eigenvalues of } n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta,\psi)>\underline{\lambda}_J \text{ for all } (\theta,\psi)\in\mathcal{N}_0(\delta)\right\}\to 1. \tag{56}$$

To prove the convergence result of (56), we first appeal to the Uniform Law of Large Numbers (ULLN) in White and Domowitz (1984): the sample averages $g_{\mathbb{Q},n}(\theta,\psi;\mathbf{x^n},\mathbf{y^n})$, $\nabla g_{\mathbb{Q},n}(\theta,\psi;\mathbf{x^n},\mathbf{y^n})$, and $\nabla^2 g_{\mathbb{Q},n}(\theta,\psi;\mathbf{x^n},\mathbf{y^n})$ converge to their population means in probability uniformly over $\Theta\times\Psi$ due to the dominance condition and the mixing condition. It implies that

$$n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta,\psi)\to 2G_{\mathbb{Q}}(\theta,\psi)^T S_{\mathbb{Q}}^{-1}G_{\mathbb{Q}}(\theta,\psi) \ \text{ in } \mathbb{Q}_n \text{ uniformly over } (\theta,\psi). \tag{57}$$

Because of the second-order continuous differentiability of the moment function and the dominance condition, the limit on the right hand side of (57) is a continuous function of (θ,ψ) due to the Dominance Convergence Theorem. Moreover, at (θ_0,ψ_0), the limiting function in (57) is equal to $G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1}G_{\mathbb{Q}}$ which is positive definite. Thus, the uniform convergence and the continuity of the limiting function directly imply the result in (56). Therefore, under the reparametrization in (48), the result implies that for an arbitrary constant $K > 0$, the probability that the smallest eigenvalue of $\nabla^2 \widehat{J}(\hat{\theta}^{\mathbb{P}} + u/\sqrt{n}, \hat{\psi}^{\mathbb{P}} + h/\sqrt{n})$ is bigger than $\underline{\lambda}_J$ converges to one uniformly over $|u| \leq K$ and $|h| \leq K$.

It suffices to focus on the big probability set in (57).

(4) First-order conditions of GMM estimators. The GMM estimator $(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}}) = (\hat{\theta}^{\mathbb{P}}, \hat{\psi}^{\mathbb{P}}) + \frac{1}{\sqrt{n}}(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{Q}})$ satisfies the first-order condition

$$0 = \nabla \widehat{J}_{n,S_{\mathbb{Q}}}(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}}). \tag{58}$$

The constrained GMM estimator $(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}})$ with restriction $\check{\theta}^{\mathbb{Q}} = \theta$ is the minimizer of $\widehat{J}_{n,S_{\mathbb{Q}}}(\tilde{\theta}, \tilde{\psi})$ subject to the constraint $\mathcal{R}(\tilde{\theta}, \tilde{\psi}; \theta, \psi) = 0$ where

$$\mathcal{R}(\tilde{\theta}, \tilde{\psi}; \theta, \psi) \equiv \mathbf{v}\Gamma_{\Theta} \begin{pmatrix} \tilde{\theta} - \theta \\ \tilde{\psi} - \psi \end{pmatrix} = \frac{1}{\sqrt{n}} \mathbf{v}\Gamma_{\Theta} \begin{pmatrix} \tilde{u} - u \\ \tilde{h} - h \end{pmatrix} = \frac{1}{\sqrt{n}} \mathcal{R}(\tilde{u}, \tilde{h}; u, h).$$

The constraint is linear and effectively restricts the first D_f elements of the baseline parameter vector to be $\theta_1 = \mathbf{v}\theta$. Recall that θ_1 is the first D_t elements of θ. The gradient of the constraint is $\nabla \mathcal{R}(\tilde{\theta}, \tilde{\psi}) \equiv \mathbf{v}\Gamma_{\Theta}$. The first-order condition and the complementarity condition (equality constraint):

$$\nabla \widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}}) = (\mathbf{v}\Gamma_{\Theta})^T \Lambda_n, \tag{59}$$

where Λ_n is a $D_f \times 1$ vector containing the Lagrangian multipliers of constraints, and

$$\mathcal{R}(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}}; \theta, \psi) = 0. \tag{60}$$

The constrained GMM $(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}})$ depends on the parameters $u_1 = \mathbf{v}u$ or $\theta_1 = \mathbf{v}\theta$.

(5) Uniform bounds and convergence for constrained GMM estimators. Under the standard regularity conditions (especially the dominance condition), it is straightforward to show that (reparametrized) GMM estimators $(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{P}})$ are $O_p(1)$ variables. However, the (reparametrized) constrained GMM estimators $(\check{u}^{\mathbb{Q}}, \check{h}^{\mathbb{Q}})$ depend on the restriction parameter $u_1 = \mathbf{v}u$, and thus it is unclear whether the constrained GMM estimators are uniformly $O_p(1)$ variables over all u, even though $(\check{u}^{\mathbb{Q}}, \check{h}^{\mathbb{Q}})$ are $O_p(1)$ for each fixed u.

In fact, the second-order Taylor expansion of $J_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}})$ around $(\hat{\theta}^{\mathbb{Q}}, \hat{\psi}^{\mathbb{Q}})$ implies that

$$d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\} = \begin{pmatrix} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{pmatrix}^T \left[n^{-1} \nabla^2 \widehat{J}_{n,S_{\mathbb{Q}}}(\theta_u, \psi_u)\right] \begin{pmatrix} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{pmatrix} \tag{61}$$

where (θ_u, ψ_u) lies on the segment between $(\hat{\theta}^{\mathbb{P}} + \check{u}^{\mathbb{Q}}/\sqrt{n},\ \hat{\psi}^{\mathbb{P}} + \check{h}^{\mathbb{Q}}/\sqrt{n})$ and $(\hat{\theta}^{\mathbb{P}} + \hat{u}^{\mathbb{Q}}/\sqrt{n},\ \hat{\psi}^{\mathbb{P}} + \hat{h}^{\mathbb{Q}}/\sqrt{n})$.

On the big probability set of (57) in Step (3), it holds that

$$d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\} \geq \underline{\lambda}_J \left(|\check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}}|^2 + |\check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}}|^2\right). \tag{62}$$

Combining (55) and (62), it holds that for any $\epsilon > 0$, there exists a large enough constant $K > 0$ and a small enough $\underline{\lambda}_J > 0$ such that

$$\limsup_{n \to \infty} \mathbb{Q}_n \left\{|\check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}}|^2 + |\check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}}|^2 \leq \underline{\lambda}_J^{-1} D_{\mathbb{Q},n}(\mathbf{x^n}, \mathbf{y^n})(|\hat{u}^{\mathbb{Q}}|^2 + K^2),\ \forall\ |u| \leq K\right\} < \epsilon$$

where $D_{\mathbb{Q},n}(\mathbf{x^n}, \mathbf{y^n})$ is defined in (54). This result is crucial since it implies that the constrained

GMM estimators $(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}})$ converges to (θ_0, ψ_0) with the rate of $\sqrt{n}$ in probability uniformly over $|u| \leq K$ for large enough constant K. Therefore, according to the ULLN result of (57), it follows that for large enough K

$$n^{-1}\nabla^2 \widehat{J}_{n,S_{\mathbb{Q}}}(\theta_u, \psi_u) \to G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}, \text{ uniformly over } |u| \leq K \text{ in } \mathbb{Q}_n. \tag{63}$$

And thus, combining the second-order Taylor expansion relationship (61) and the asymptotic result of (63), it implies that

$$d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\} = \begin{pmatrix} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{pmatrix}^T G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}} \begin{pmatrix} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{pmatrix} + o_{p,K}(1), \tag{64}$$

where the term $o_{p,K}(1)$ converges to zero uniformly over $|u| \leq K$ in probability.

(6) A normal approximation of weighting posterior distributions in the integral. As a result of the uniform bound on the over-fitting gap $d_{S_{\mathbb{Q}}}\{\mathbf{v}\theta; \mathbf{x^n}, \mathbf{y^n}\}$, the approximating error of replacing $\pi_{\mathbb{P}}(\theta|\mathbf{x^n})$ by the normal density for $N(\hat{\theta}^{\mathbb{P}}, n^{-1}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1})$, denoted by $\varphi_{\mathbb{P},n}(\theta)$, can be bounded as follows

$$\begin{aligned} &|\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) - \int d_{S_{\mathbb{Q}}}\{\mathbf{v}\theta; \mathbf{x^n}, \mathbf{y^n}\}\varphi_{\mathbb{P},n}(\theta)\mathrm{d}\theta| \\ &\qquad \leq \int d_{S_{\mathbb{Q}}}\{\mathbf{v}\theta; \mathbf{x^n}, \mathbf{y^n}\}|\pi_{\mathbb{P}}(\theta|\mathbf{x^n}) - \varphi_{\mathbb{P},n}(\theta)|\mathrm{d}\theta. \end{aligned} \tag{65}$$

Change of variable $\theta = \hat{\theta}^{\mathbb{P}} + \frac{u}{\sqrt{n}}$, the inequality (65) can be rewritten as

$$\begin{aligned} &|\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) - \int d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u| \\ &\qquad \leq \int d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}|\pi_{\mathbb{P}}(\hat{\theta}^{\mathbb{P}} + u/\sqrt{n}|\mathbf{x^n}) - \varphi_{\mathbb{P}}(u)|\mathrm{d}u. \end{aligned} \tag{66}$$

From the inequality above (66) and (55), it follows that

$$\begin{aligned} &|\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) - \int d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u| \\ &\qquad \leq D_{\mathbb{Q},n}(\mathbf{x^n}, \mathbf{y^n}) \int |u - \hat{u}^{\mathbb{Q}}|^2 |\pi_{\mathbb{P}}(\hat{\theta}^{\mathbb{P}} + u/\sqrt{n}|\mathbf{x^n}) - \varphi_{\mathbb{P}}(u)|\mathrm{d}u. \end{aligned} \tag{67}$$

Convergence in relative entropy implies convergence in total variation distance due to the Pinsker's inequality (see, e.g. Pinsker, 1964). Thus, Proposition 4 and the intermediate convergence result (45) imply that

$$\int |u - \hat{u}^{\mathbb{Q}}|^2 |\pi_{\mathbb{P}}(\hat{\theta}^{\mathbb{P}} + u/\sqrt{n}|\mathbf{x^n}) - \varphi_{\mathbb{P}}(u)|\mathrm{d}u. = o_p(1). \tag{68}$$

Therefore, because $D_{\mathbb{Q},n}\{\mathbf{x^n}, \mathbf{y^n}\}$ defined in (55) is $O_p(1)$, it holds that

$$\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) = \int d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u + o_p(1). \tag{69}$$

Thus, as far as the asymptotic properties of $\varrho_o^f(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n})$ are concerned, we need to only focus on the integral in (69).

(7) A Wald-type approximation for the over-fitting gap $d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}$. Let's define a Wald-type distance between u_1 and $\hat{u}_1^{\mathbb{Q}}$:

$$w_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\} = \left(\begin{array}{c} \hat{u}^{\mathbb{Q}} - u \\ \hat{h}^{\mathbb{Q}} - h \end{array}\right)^T \Gamma_{\Theta}^T \mathbf{v}^T \left\{\mathbf{v}\Gamma_{\Theta}\left[G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right]^{-1}\Gamma_{\Theta}^T\mathbf{v}^T\right\}^{-1}\mathbf{v}\Gamma_{\Theta}\left(\begin{array}{c} \hat{u}^{\mathbb{Q}} - u \\ \hat{h}^{\mathbb{Q}} - h \end{array}\right)$$

Based on the definition of $\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)$ in (9), the Wald-type distance above can be rewritten as

$$\begin{aligned} w_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\} &= \left(\hat{u}^{\mathbb{Q}} - u\right)^T \mathbf{v}^T\left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1}\mathbf{v}\left(\hat{u}^{\mathbb{Q}} - u\right) && (70) \\ &= \mathcal{R}(\hat{u}^{\mathbb{Q}},\hat{h}^{\mathbb{Q}};u,h)^T\Gamma_{\Theta}^T\mathbf{v}^T\left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1}\mathbf{v}\Gamma_{\Theta}\mathcal{R}(\hat{u}^{\mathbb{Q}},\hat{h}^{\mathbb{Q}};u,h). && (71) \end{aligned}$$

Our idea is to approximate $d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}$ by the Wald-type distance $w_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}$. More precisely, we shall show that

$$\int d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u = \int w_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u + o_p(1). \tag{72}$$

Because of the gaussian tail of $\varphi_{\mathbb{P}}(u)$ and the quadratic bounds of $d_{S_{\mathbb{Q}}}$ and $w_{S_{\mathbb{Q}}}$, it suffices to show that for any large enough constant K,

$$\int_{|u|\leq K} d_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u = \int_{|u|\leq K} w_{S_{\mathbb{Q}}}\{\mathbf{v}u;\mathbf{x^n},\mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u + o_p(1). \tag{73}$$

(8) Proof of the Wald-type approximation in (73). The starting point of the proof of Wald-type approximations goes back to the second-order Taylor expansion of $J_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}})$ around $(\hat{\theta}^{\mathbb{Q}},\hat{\psi}^{\mathbb{Q}})$ and its large-sample approximation in (64). Now we represent $\left(\begin{array}{c} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{array}\right)$ in the approximation relationship (64) by $\mathcal{R}(\hat{u}^{\mathbb{Q}},\hat{h}^{\mathbb{Q}};u,h)$. The most direct connection between them is obviously based on the first-order Taylor expansion of $\mathcal{R}(\hat{u}^{\mathbb{Q}},\hat{h}^{\mathbb{Q}};u,h)$ around $(\check{u}^{\mathbb{Q}},\check{h}^{\mathbb{Q}})$:

$$\mathcal{R}(\hat{u}^{\mathbb{Q}},\hat{h}^{\mathbb{Q}};u,h) = -\mathbf{v}\Gamma_{\Theta}\left(\begin{array}{c} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{array}\right) \tag{74}$$

However, $\mathbf{v}\Gamma_{\Theta}$ is not invertible, and thus we need to figure out another strategy. Basically, it is to first represent $\left(\begin{array}{c} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{array}\right)$ by $\nabla\widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}})$, and then we represent $\nabla\widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}})$ by the Lagrangian multiplier Λ, and lastly we represent the Lagrangian multiplier Λ_n by $\mathcal{R}(\hat{u}^{\mathbb{Q}},\hat{h}^{\mathbb{Q}};u,h)$.

Again, we use the first-order Taylor expansion of $\nabla\widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}})$ around $(\hat{\theta}^{\mathbb{Q}},\hat{\psi}^{\mathbb{Q}})$:

$$\nabla\widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}}) = n^{-1}\nabla^2\widehat{J}_{n,S_{\mathbb{Q}}}(\theta'_u,\psi'_u)\left(\begin{array}{c} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{array}\right) \tag{75}$$

where (θ'_u,ψ'_u) lies on the segment between $(\hat{\theta}^{\mathbb{P}}+\check{u}^{\mathbb{Q}}/\sqrt{n},\ \hat{\psi}^{\mathbb{P}}+\check{h}^{\mathbb{Q}}/\sqrt{n})$ and $(\hat{\theta}^{\mathbb{P}}+\hat{u}^{\mathbb{Q}}/\sqrt{n},\ \hat{\psi}^{\mathbb{P}}+\hat{h}^{\mathbb{Q}}/\sqrt{n})$. Thus,

$$\nabla\widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}}) = G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\left(\begin{array}{c} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{array}\right) + o_{p,K}(1), \tag{76}$$

where the term $o_{p,K}(1)$ converges to zero uniformly over $|u|\leq K$ in probability. Thus, it holds that

$$\left(\begin{array}{c} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{array}\right) = \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1}\nabla\widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}},\check{\psi}^{\mathbb{Q}}) + o_{p,K}(1). \tag{77}$$

Now we represent $\nabla \widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}})$ by the Lagrangian multiplier Λ_n using the first-order condition of constrained GMM estimators, which is Equation (59):

$$\nabla \widehat{J}_{n,S_{\mathbb{Q}}}(\check{\theta}^{\mathbb{Q}}, \check{\psi}^{\mathbb{Q}}) = (\mathbf{v}\Gamma_\Theta)^T \Lambda_n. \tag{78}$$

Plugging the equation above into (74), it follows that

$$\mathcal{R}(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{Q}}; u, h) = -\mathbf{v}\Gamma_\Theta \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1} (\mathbf{v}\Gamma_\Theta)^T \Lambda_n + o_{p,K}(1). \tag{79}$$

Thus,

$$\Lambda_n = -\left\{\mathbf{v}\Gamma_\Theta \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1} (\mathbf{v}\Gamma_\Theta)^T\right\}^{-1} \mathcal{R}(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{Q}}; u, h) + o_{p,K}(1). \tag{80}$$

Combining (77), (78), and (80), it leads to

$$\begin{aligned}
&\begin{pmatrix} \check{u}^{\mathbb{Q}} - \hat{u}^{\mathbb{Q}} \\ \check{h}^{\mathbb{Q}} - \hat{h}^{\mathbb{Q}} \end{pmatrix} \\
&= \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1} \Gamma_\Theta^T \mathbf{v}^T \left[\mathbf{v}\Gamma_\Theta \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1} \Gamma_\Theta^T \mathbf{v}^T\right]^{-1} \mathcal{R}(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{Q}}; u, h) + o_{p,K}(1).
\end{aligned}$$

Now, we plug the relationship above into (64), and we obtain the following approximation:

$$\begin{aligned}
&d_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\} \\
&\quad = \mathcal{R}(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{Q}}; u, h)^T \Gamma_\Theta^T \mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\Gamma_\Theta \mathcal{R}(\hat{u}^{\mathbb{Q}}, \hat{h}^{\mathbb{Q}}; u, h) + o_{p,K}(1).
\end{aligned} \tag{81}$$

Comparing (71) and (81), we have shown what we promised earlier in Equation (73). In other words, we have established the large-sample relationship in (72). In the end, we derive the asymptotic distribution of $\int w_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u$ as follows.

(9) Asymptotic distribution of $\int w_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\}\varphi_{\mathbb{P}}(u)\mathbf{d}u$. We consider the decomposition

$$w_{S_0}\{\mathbf{v}u; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\} = u^T\mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}u \tag{82}$$

$$+ 2(\hat{u}^{\mathbb{Q}})^T\mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}u \tag{83}$$

$$+ (\hat{u}^{\mathbb{Q}})^T\mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\hat{u}^{\mathbb{Q}}. \tag{84}$$

The integral of term (83) over the Gaussian density $\varphi_{\mathbb{P}}(u)$ is zero, while the integral of term (84) is just itself. Lastly, the integral of term (82) over the Gaussian density is

$$\int \mathbf{tr}\left\{\left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}uu^T\mathbf{v}^T\right\}\varphi_{\mathbb{P}}(u)\mathrm{d}u = \varrho^{\mathbf{v}}(\theta_0|\psi_0). \tag{85}$$

Thus, the integral can be rewritten as

$$\int w_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x}^{\mathbf{n}}, \mathbf{y}^{\mathbf{n}}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u = \varrho^{\mathbf{v}}(\theta_0|\psi_0) + (\hat{u}^{\mathbb{Q}})^T\mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\hat{u}^{\mathbb{Q}}. \tag{86}$$

To derive the asymptotic approximation of $(\hat{u}^{\mathbb{Q}})^T\mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\hat{u}^{\mathbb{Q}}$, we appeal to the standard large-sample approximations of GMM in the literature; see, for example, Hansen (1982, Theorem 3.1) and Newey (1985a, Lemma 1 and Theorem 3). More precisely, we have

$$\mathbf{v}\hat{u}^{\mathbb{Q}} = -\mathbf{v}\left[\Gamma_\Theta \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1} G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} - \left(G_{\mathbb{P}}^T S_{\mathbb{P}}^{-1} G_{\mathbb{P}}\right)^{-1} G_{\mathbb{P}} S_{\mathbb{P}}^{-1}\Gamma_{\mathbb{P}}\right] \sqrt{n}\hat{g}_{\mathbb{Q},n}(\theta_0, \psi_0) + o_p(1),$$

where $\Gamma_{\mathbb{P}} = \left[I_{D_{\mathbb{P}}}, O_{D_{\mathbb{P}}\times(D_{\mathbb{Q}}-D_{\mathbb{P}})}\right]$ and $\Gamma_{\Theta} = \left[I_{D_f}, O_{D_f\times(D_\Theta-D_f)}\right]$.

Thus, $\hat{u}^{\mathbb{Q}}$ is asymptotically normally distributed

$$\operatorname*{wlim}_{n\to\infty} \mathbf{v}\hat{u}^{\mathbb{Q}} = \mathbf{v}\Upsilon_{\mathbb{Q}} Z$$

where $Z \sim N(0, I_{D_{\mathbb{Q}}})$ and

$$\Upsilon_{\mathbb{Q}} \equiv -\left[\Gamma_\Theta \left(G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} G_{\mathbb{Q}}\right)^{-1} G_{\mathbb{Q}}^T S_{\mathbb{Q}}^{-1} - \left(G_{\mathbb{P}}^T S_{\mathbb{P}}^{-1} G_{\mathbb{P}}\right)^{-1} G_{\mathbb{P}}^T S_{\mathbb{P}}^{-1} \Gamma_{\mathbb{P}}\right] S_{\mathbb{Q}}^{1/2}. \tag{87}$$

The Continuous Mapping Theorem implies that

$$\operatorname*{wlim}_{n\to\infty} (\hat{u}^{\mathbb{Q}})^T \mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\hat{u}^{\mathbb{Q}} = Z^T \Upsilon_{\mathbb{Q}}^T \mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\Upsilon_{\mathbb{Q}} Z. \tag{88}$$

Consider the Singular Value Decomposition (SVD):

$$\Upsilon_{\mathbb{Q}}^T \mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1/2} = \mathbf{U}\mathbf{\Sigma}\widetilde{\mathbf{U}}^T,$$

where $\mathbf{U}$ is a $D_{\mathbb{Q}} \times D_{\mathbb{Q}}$ orthogonal matrix, $\widetilde{\mathbf{U}}$ is a $D_f \times D_f$ orthogonal matrix, and $\mathbf{\Sigma}$ is a $D_{\mathbb{Q}} \times D_f$ diagonal matrix with singular values on the diagonal line. Plugging back into (88), it implies the following convergence:

$$\begin{aligned}\operatorname*{wlim}_{n\to\infty} (\hat{u}^{\mathbb{Q}})^T \mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1} \mathbf{v}\hat{u}^{\mathbb{Q}} &= \left(\mathbf{U}^T Z\right)^T \mathbf{\Sigma}\mathbf{\Sigma}^T \mathbf{U}^T Z \\ &= \sigma_1 z_1^2 + \cdots + \sigma_{D_f} z_{D_f}^2,\end{aligned} \tag{89}$$

where $\mathbf{U}Z = (z_1, \cdots, z_{D_f}, z_{D_f+1}, \cdots, z_{D_\Theta})^T$ contains D_Θ i.i.d. standard normal random variables, and $\sigma_1, \cdots, \sigma_{D_f}$ are the nonzero diagonal elements of $\mathbf{\Sigma}\mathbf{\Sigma}^T$. In fact, $\sigma_1, \cdots, \sigma_{D_f}$ are actually the eigenvalues of the matrix

$$\widetilde{\mathbf{U}}\mathbf{\Sigma}^T\mathbf{\Sigma}\widetilde{\mathbf{U}}^T = \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1/2} \mathbf{v}\Upsilon_{\mathbb{Q}}\Upsilon_{\mathbb{Q}}^T\mathbf{v}^T \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1/2} \tag{90}$$

From the definition of $\Upsilon_{\mathbb{Q}}$ in (87), it follows that

$$\Upsilon_{\mathbb{Q}}\Upsilon_{\mathbb{Q}}^T = \mathbf{I}_{\mathbb{P}}(\theta_0)^{-1} - \mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}. \tag{91}$$

In deriving the equality of (91), there are two relationships that worth mentioning on top of straightforward matrix algebra:

$$\Gamma_{\mathbb{P}} S_{\mathbb{Q}} \Gamma_{\mathbb{P}}^T = S_{\mathbb{P}} \quad \text{and} \quad \Gamma_{\mathbb{P}} G_{\mathbb{Q}} = [G_{\mathbb{P}}, O_{D_{\mathbb{P}}\times D_\Psi}] = [G_{\mathbb{P}}, O_{D_{\mathbb{P}}\times D_\Psi}]\,\Gamma_\Theta^T\Gamma_\Theta. \tag{92}$$

Plugging (91) back into (90), we know that $\sigma_1, \cdots, \sigma_{D_f}$ are actually the eigenvalues of the matrix

$$\begin{aligned}\widetilde{\mathbf{U}}\mathbf{\Sigma}^T\mathbf{\Sigma}\widetilde{\mathbf{U}}^T &= \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1/2} \left[\mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}^T\right] \left[\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right]^{-1/2} - I_{D_f} \\ &= \Pi_0(\mathbf{v}) - I_{D_f},\end{aligned}$$

where $\Pi_0(\mathbf{v})$ is the (Fisher) information matrix defined in (13). Thus, the eigenvalues are

$$\sigma_i = \lambda_i(\mathbf{v}) - 1, \quad \text{for all } i = 1, \cdots, D_f. \tag{93}$$

Here, the $\lambda_i(\mathbf{v})$'s are eigenvalues of $\Pi_0(\mathbf{v})$ as in Proposition 1. Therefore, according to (86), (89), and (93), it follows that

$$\underset{n\to\infty}{\text{wlim}} \int w_{S_{\mathbb{Q}}}\{\mathbf{v}u; \mathbf{x^n}, \mathbf{y^n}\}\varphi_{\mathbb{P}}(u)\mathrm{d}u = \varrho^{\mathbf{v}}(\theta_0|\psi_0) + \sum_i^{D_f} \left[\lambda_i(\mathbf{v}) - 1\right] \chi^2_{1,i}, \tag{94}$$

where $\chi^2_{1,i}$'s are i.i.d. chi-squared random variables with 1 degree of freedom.

B.3 Proof of The Corollaries

The result of Corollary 1 follows immediately from Theorem 1. The result of Corollary 2 follows from Theorem 1 of Nakagawa (2005). More precisely, the Laplace-Stieltjes transform of the cumulative distribution function of $\text{wlim}_{n\to\infty} \varrho^f_o(\theta_0|\psi_0, \mathbf{x^n}, \mathbf{y^n}) = \varrho^{\mathbf{v}}(\theta_0|\psi_0) + \sum_{i=1}^{D_f} \left[\lambda_i(\mathbf{v}) - 1\right] \chi^2_{1,i}$ is

$$\mathcal{M}(z) \equiv \mathbb{E}_{\mathbb{Q}} e^{-z\left\{\varrho^{\mathbf{v}}(\theta_0|\psi_0) + \sum_{i=1}^{D_f}[\lambda_i(\mathbf{v})-1]\chi^2_{1,i}\right\}} = e^{-z\varrho^{\mathbf{v}}(\theta_0|\psi_0)} \prod_{i=1}^{D_f} \left[1 + 2z\left(\lambda_i(\mathbf{v}) - 1\right)\right]^{-1/2}. \tag{95}$$

Let $\mathcal{R}z$ denote the real part of z. Thus, the abscissa of convergence of $\mathcal{M}(z)$ is equal to $-\frac{1}{2[\lambda_1(\mathbf{v})-1]}$ where $\lambda_1(\mathbf{v})$ is the largest eigenvalue; that is, when $\mathcal{R}z > -\frac{1}{2[\lambda_1(\mathbf{v})-1]}$, the transform $\mathcal{M}(z)$ converges, and when $\mathcal{R}z < -\frac{1}{2[\lambda_1(\mathbf{v})-1]}$, the transform $\mathcal{M}(z)$ diverges. Therefore, according to Theorem 1 of Nakagawa (2005), the tail probability has the convergence property stated in the corollary.

C Proof of Propositions on Fisher Model Fragility

C.1 Proof of Proposition 1

It is straightforward from the fact that

$$\begin{aligned}
&\mathbf{tr}\left[\left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1}\left(\mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}^T\right)\right] \\
&\quad = \mathbf{tr}\left[\left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1/2}\left(\mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{v}^T\right)\left(\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T\right)^{-1/2}\right]
\end{aligned}$$

and the fact that trace of a symmetric matrix is equal to the sum of its all eigenvalues. Because $g_{\mathbb{P}}(\theta; \mathbf{x_t})$ is part of $g_{\mathbb{Q}}(\theta, \psi; \mathbf{x_t}, \mathbf{y_t})$, according to Hansen (1982), the asymptotic covariance matrices satisfy $\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1} \le \mathbf{I}_{\mathbb{P}}(\theta_0|\psi_0)^{-1}$. Thus, for any full-rank matrix $\mathbf{v}$, it holds that $\mathbf{v}\mathbf{I}_{\mathbb{Q}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T \le \mathbf{v}\mathbf{I}_{\mathbb{P}}(\theta_0|\psi_0)^{-1}\mathbf{v}^T$. As a result, the smallest eigenvalue is not less than one.

C.2 Proof of Proposition 2

If we define $\mathbf{u} = \mathbf{I}_{\mathbb{Q}}(\theta_0)^{-1/2}\mathbf{v}$, we can rewrite the $\varrho^D_a(\theta_0)$ as

$$\begin{aligned}
\varrho^D_a(\theta_0) &= \max_{\mathbf{u}\in\mathbb{R}^{D_\Theta\times D}, \mathbf{Rank}(\mathbf{u})=D} \mathbf{tr}\left[\left(\mathbf{u}^T\mathbf{u}\right)^{-1}\left(\mathbf{u}^T\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{u}\right)\right] \\
&= \max_{\mathbf{u}\in\mathbb{R}^{D_\Theta\times D}, \mathbf{Rank}(\mathbf{u})=D} \mathbf{tr}\left[\mathbf{u}\left(\mathbf{u}^T\mathbf{u}\right)^{-1}\mathbf{u}^T\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\right]
\end{aligned}$$

The linear operator $\mathfrak{P}_{\mathbf{u}} \equiv \mathbf{u}\left(\mathbf{u}^T\mathbf{u}\right)^{-1}\mathbf{u}^T$ is the projection operator onto the subspace spanned by the column vectors of $\mathbf{u}$. Therefore, we have

$$\varrho_a^D(\theta_0) = \max_{\mathbf{u}\in\mathbb{R}^{D_\Theta\times D}, \mathbf{Rank}(\mathbf{u})=D} \mathbf{tr}\left[\mathfrak{P}_{\mathbf{u}}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\right].$$

The projection operator can be equivalently expressed in terms of the orthonormal column vectors lying in the subspace spanned by $\mathbf{u}$. Thus, without loss of any generality, we can assume that the column vectors of $\mathbf{u}$ are orthonormal vectors, i.e. $\mathbf{u}^T\mathbf{u}$ is a D-dimensional identity matrix. Therefore,

$$\begin{aligned}
\varrho_a^D(\theta_0) &= \max_{\mathbf{u}\in\mathbb{R}^{D_\theta\times D}, \mathbf{Rank}(\mathbf{u})=D, \mathbf{u}^T\mathbf{u}=I} \mathbf{tr}\left[\mathbf{u}\mathbf{u}^T\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\right] \\
&= \max_{\mathbf{u}\in\mathbb{R}^{D_\theta\times D}, \mathbf{Rank}(\mathbf{u})=D, \mathbf{u}^T\mathbf{u}=I} \mathbf{tr}\left[\mathbf{u}^T\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{u}\right] \\
&= \max_{\mathbf{u}\in\mathbb{R}^{D_\theta\times D}, \mathbf{Rank}(\mathbf{u})=D, \mathbf{u}^T\mathbf{u}=I} \sum_{i=1}^{D} u_i^T\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}u_i \\
&= \lambda_1 + \lambda_2 + \cdots + \lambda_D.
\end{aligned}$$

The argmax matrix is $\mathbf{u}^* = [e_1^*, e_2^*, \cdots, e_D^*]$ whose column vectors are the corresponding eigenvectors. Thus, correspondingly, the worst-case matrix is $\mathbf{v}^* = [v_1^*, v_2^*, \cdots, v_D^*]$ with $v_i^* = \mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}e_i^*$. Moreover, for $i = 1, \cdots, D$, it holds that

$$\lambda_i = \frac{(v_i^*)^T\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}v_i^*}{(v_i^*)^T\mathbf{I}_{\mathbb{Q}}(\theta_0)^{-1}v_i^*}. \tag{96}$$

We have completed the proof. Furthermore, the eigen problem above is equivalent to the one

$$\lambda_i = \frac{(\hat{v}_i^*)^T\mathbf{I}_{\mathbb{Q}}(\theta_0)\hat{v}_i^*}{(\hat{v}_i^*)^T\mathbf{I}_{\mathbb{P}}(\theta_0)\hat{v}_i^*}, \tag{97}$$

where $\hat{v}_i^* = \mathbf{I}_{\mathbb{P}}(\theta_0)^{-1/2}\hat{e}_i^*$ with $\hat{e}_i^*$ to be the corresponding eigenvector of $\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1/2}\mathbf{I}_{\mathbb{Q}}(\theta_0)\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1/2}$. That is, the vector $\hat{v}_i^*$ corresponds to v_i^*.

C.3 Proof of Proposition 3

From Proposition 2, it holds that

$$\varrho^{D_1}(\theta_0|\psi_0) = \lambda_1 + \cdots + \lambda_{D_1}, \text{ and } \varrho^{D_2}(\theta_0|\psi_0) = \lambda_1 + \cdots + \lambda_{D_2}. \tag{98}$$

As the eigenvalues are nonnegative, it must hold that $\varrho^{D_1}(\theta_0|\psi_0) \leq \varrho^{D_2}(\theta_0|\psi_0)$ when $D_1 \leq D_2$.

D Derivations of The Disaster Risk Model

We first show how to derive the Euler equation, and then we show how to obtain the Fisher fragility measure $\varrho(p, \lambda)$.

D.1 The Euler Equation

The total return of market equity from t to $t+1$ is $e^{r_{M,t+1}}$ which is unknown at t, and the total interest gain of risk-free bond is $e^{r_{f,t}}$ which is known at t. Thus, the excess log return of equity is $r_{t+1} = r_{M,t+1} - r_{f,t}$. The state-price density is $\Lambda_t = \delta_{\mathrm{D}}^t c_t^{-\gamma_{\mathrm{D}}}$, and the inter-temporal marginal rate of substitution is $\Lambda_{t+1}/\Lambda_t = \delta_{\mathrm{D}} e^{-\gamma_{\mathrm{D}} g_{t+1}}$. The Euler equations for risk-free rate and the market equity return are

$$1 = \mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t} e^{r_{M,t+1}}\right] \quad \text{and} \quad e^{-r_{f,t}} = \mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t}\right]. \tag{99}$$

Thus, we can obtain the Euler equation for the excess log return:

$$\mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t}\right] = \mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t} e^{r_{t+1}}\right]. \tag{100}$$

The left-hand side of (100) can be computed as

$$\mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t}\right] = \mathbb{E}_t\left[e^{-\gamma_{\mathrm{D}} g_{t+1}}\right] = (1-p)e^{-\gamma_{\mathrm{D}}\mu + \frac{1}{2}\gamma_{\mathrm{D}}^2\sigma^2} + p\lambda\frac{e^{\gamma_{\mathrm{D}}\underline{v}}}{\lambda - \gamma_{\mathrm{D}}},$$

and the right-hand side of (100) can be computed as

$$\mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t} e^{r_{t+1}}\right] = \mathbb{E}_t\left[e^{-\gamma_{\mathrm{D}} g_{t+1} + r_{t+1}}\right] = (1-p)e^{-\gamma_{\mathrm{D}}\mu + \eta + \frac{1}{2}(\gamma_{\mathrm{D}}^2\sigma^2 + \tau^2 - 2\gamma_{\mathrm{D}}\rho\sigma\tau)} + p\lambda\frac{e^{\frac{\nu^2}{2} + (\gamma_{\mathrm{D}} - b)\underline{v}}}{\lambda + b - \gamma_{\mathrm{D}}}$$

Thus, the Euler equation (100) can be rewritten as

$$(1-p)e^{-\gamma_{\mathrm{D}}\mu + \frac{1}{2}\gamma_{\mathrm{D}}^2\sigma^2}\left[e^{\eta + \frac{1}{2}\tau^2 - \gamma_{\mathrm{D}}\rho\sigma\tau} - 1\right] = p\Delta(\lambda), \tag{101}$$

where

$$\Delta(\lambda) = \lambda\left(\frac{e^{\gamma_{\mathrm{D}}\underline{v}}}{\lambda - \gamma_{\mathrm{D}}} - \frac{e^{\frac{\nu^2}{2} + (\gamma_{\mathrm{D}} - b)\underline{v}}}{\lambda + b - \gamma_{\mathrm{D}}}\right).$$

Rearranging terms in (101), it leads to the final Euler equation in (29). Using the Taylor expansion, we have the following approximation:

$$e^{\eta + \frac{1}{2}\tau^2 - \gamma_{\mathrm{D}}\rho\sigma\tau} - 1 \approx \eta + \frac{1}{2}\tau^2 - \gamma_{\mathrm{D}}\rho\sigma\tau. \tag{102}$$

Combining (101) and the approximation in (102), we have finished proving the approximated Euler equation in (30).

D.2 Fisher fragility measure

The joint probability density for rare disasters (z, v) in the baseline model is

$$f_{\mathbb{P}}(z, \tilde{v}|p, \lambda) = p^z(1-p)^{1-z}\delta(v)^{1-z}\left[\mathbf{1}\{v > \underline{v}\}\lambda\exp\{-\lambda(v - \underline{v})\}\right]^z, . \tag{103}$$

where $\delta(\cdot)$ is the dirac delta function. The Fisher information matrix is

$$\mathbf{I}_{\mathbb{P}}(p, \lambda) = \begin{bmatrix} \frac{1}{p(1-p)} & 0 \\ 0 & \frac{p}{\lambda^2} \end{bmatrix}. \tag{104}$$

Next, the probability density function $f_{\mathbb{Q}}(z,v,r,u|\theta)$ for the structural model is

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

where

$$
\eta(\theta,\phi) \equiv \gamma_{\mathrm{D}}\rho\sigma\tau - \frac{\tau^2}{2} + \ln\left[1 + e^{\gamma_{\mathrm{D}}\mu - \frac{\gamma_{\mathrm{D}}^2\sigma^2}{2}} \lambda \left(\frac{e^{\gamma_{\mathrm{D}}\underline{v}}}{\lambda-\gamma_{\mathrm{D}}} - e^{\frac{1}{2}\nu^2} \frac{e^{(\gamma_{\mathrm{D}}-b)\underline{v}}}{\lambda+b-\gamma_{\mathrm{D}}} \right) \frac{p}{1-p} \right]. \tag{105}
$$

We can derive the simple intuitive closed-form approximation for the fragility measure in (32), if we consider the approximated Euler equation. More precisely, we consider the following approximation:

$$
\eta(\theta,\phi) \approx \gamma_{\mathrm{D}}\rho\sigma\tau - \frac{\tau^2}{2} + e^{\gamma_{\mathrm{D}}\mu - \frac{\gamma_{\mathrm{D}}^2\sigma^2}{2}} \lambda \left(\frac{e^{\gamma_{\mathrm{D}}\underline{v}}}{\lambda-\gamma_{\mathrm{D}}} - e^{\frac{1}{2}\nu^2} \frac{e^{(\gamma_{\mathrm{D}}-b)\underline{v}}}{\lambda+b-\gamma_{\mathrm{D}}} \right) \frac{p}{1-p}, \tag{106}
$$

Then, using the notation introduced by (31) and (33), we can express the Fisher information for (p,λ) under the full structural model as

$$
\mathbf{I}_{\mathbb{Q}}(p,\lambda) \approx \begin{bmatrix} \frac{1}{p(1-p)} + \frac{\Delta(\lambda)^2}{(1-\rho^2)\tau^2} \frac{e^{2\gamma_{\mathrm{D}}\mu-\gamma_{\mathrm{D}}^2\sigma^2}}{(1-p)^3} & \frac{p}{(1-\rho^2)\tau^2} \frac{e^{2\gamma_{\mathrm{D}}\mu-\gamma_{\mathrm{D}}^2\sigma^2}}{(1-p)^2} \Delta(\lambda)\dot{\Delta}(\lambda) \\ \frac{p}{(1-\rho^2)\tau^2} \frac{e^{2\gamma_{\mathrm{D}}\mu-\gamma_{\mathrm{D}}^2\sigma^2}}{(1-p)^2} \Delta(\lambda)\dot{\Delta}(\lambda) & \frac{p}{\lambda^2} + \frac{\dot{\Delta}(\lambda)^2}{(1-\rho^2)\tau^2} e^{2\gamma_{\mathrm{D}}\mu-\gamma_{\mathrm{D}}^2\sigma^2} \frac{p^2}{1-p} \end{bmatrix}. \tag{107}
$$

Following Proposition 2, the worst-case Fisher fragility is the largest eigenvalue of the matrix $\Pi_0(I_{D_\Theta}) \equiv \mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1}\mathbf{I}_{\mathbb{Q}}(\theta_0)^{1/2}$. Important for simplifying the calculation, it is also the largest eigenvalue of $\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1/2}\mathbf{I}_{\mathbb{Q}}(\theta_0)\mathbf{I}_{\mathbb{P}}(\theta_0)^{-1/2}$. In this case, the eigenvalues and eigenvectors are available in closed form. This gives us the formula for $\varrho(p,\lambda)$ and $\varrho^1(p,\lambda)$ in (32). The minimum Fisher fragility in this case is 1, which is obtained in the direction along the deterministic cross-equation restriction.