Semi-parametric estimation of nonlinear rational expectations models with recursive preferences

Draft: August 2019

Abstract

We propose a semi-parametric estimator for models with Epstein-Zin preferences that deploys the entire structural information, thereby avoiding unjustified constraints. Specifically, we leave the parametrization of the Epstein-Zin stochastic discount factor intact, and exploit that stationary recursive preferences define a contraction mapping with respect to the continuation utility to obtain a sequence for the continuation utility by repeated substitution into a Nadaraya-Watson estimator. We estimate the preference parameters by minimizing a dynamic restriction on the squared pricing errors using monthly U.S. data from 1952m2 to 2019m1, and inspect the pricing errors. The estimate for the parameter of risk aversion is lower than found elsewhere in the econometric literature, and corresponds to plausible values suggested by economic theory. We evaluate the performance of our estimator with a Monte Carlo experiment. Our findings highlight the importance of the entire structural information for the model estimation.

JEL Classification: C14, C36, E44, G12

1 Introduction

Recursive preferences become increasingly popular in macroeconomics and financial economics. Macroeconomists and, in particular, financial economists increasingly build models that feature recursive preferences to match, among others, risk premia, patterns in return predictability, and time-varying Sharpe ratios.¹ Next to that, recursive preferences are used in quantitative macroeconomic models to evaluate the consequences of uncertainty shocks induced by, for instance, policymakers. Recursive preferences express the current utility level of a (representative) agent in terms of an time aggregator over current consumption and a discounted value of a risk aggregator of future consumption. The Epstein-Zin functional, for instance, features a constant elasticity of substitution (CES) aggregator over consumption and the discounted risk aggregator, where the risk aggregator is a measure of certainty equivalence of the continuation utility, i.e. the value attached to the future consumption stream.² Epstein-Zin preferences are typically used because they allow for the separation between the elasticity of intertemporal substitution and relative risk aversion, which both are represented by a single separate parameter. Under the classic Constant Relative Risk Aversion (CRRA) utility framework, the elasticity of intertemporal substitution are each others reciprocal. Epstein-Zin preferences break this link and, thus, imply that the willingness to substitute consumption across states of nature (risk aversion) and the willingness to substitute consumption across time (intertemporal substitution) are not as intimately linked as the CRRA utility framework suggests.

¹The recent article by Cochrane (2017) nests a good overview of these studies.
²Other forms of recursive preferences are discussed by Backus et al. (2004).
Recursive preferences are, however, difficult to implement in economic models. Models featuring such preferences are difficult to solve due to the high degree of non-linearities innate in these models (e.g. see Caldara et al., 2012; Pohl et al., 2018b). Next to that, the parameters governing the elasticity of intertemporal substitution and risk aversion are notoriously difficult to estimate and, consequently, models featuring recursive preferences are often calibrated with little econometric guidance. We attempt to contribute to solving the latter problem by proposing a novel estimator.

We propose a novel estimator for Epstein-Zin preferences that circumvents three identification issues in the current literature. The Epstein-Zin stochastic discount factor is a function of consumption growth and the unobservable continuation utility. Since the continuation utility is not observable, we cannot directly estimate the parameters using established statistical methods, such as Generalized Method of Moments. There are two established ways ensure identification of the model. The first one is proposed by Epstein and Zin (1991), who show that the continuation utility can be expressed in terms of the gross return on the wealth portfolio. However, the wealth portfolio spans the complete investment universe and, therefore, has to be proxied. A second way is to specify a consumption process and its distribution – using a stochastic endowment or by modeling a complete macroeconomy – to model the evolution of the continuation utility (e.g. see Bansal et al., 2010). In both these approaches, however, model specification tests are obfuscated. In the former approach, the model might be rejected because the proxy is not sufficiently broad so that it does not capture the wealth portfolio. In the latter approach, the utility model might be rejected because the assumed macroeconomic structure is an inadequate representation of reality. A third issue relates to the structure of the model. Various estimating strategies rely on linearizations of the model or linear approximations of the risk aggregator (e.g. see Chen et al., 2013; Yogo, 2004). However, the Epstein-Zin utility function and stochastic discount factor are highly non-linear functions, and so linearizations might be not at all appropriate.

To mitigate these identification issues, we leave the parametrization of the stochastic discount factor intact, and exploit the flexibility of non-parametric regression techniques to model the evolution of the continuation utility. Ergo, we do not have to proxy the return on the wealth portfolio and nor do we have to model a full-fledged macroeconomic model. We only have to make very mild assumptions on the serial dependence of the conditioning variables. A unique property of our method is that we do not have to make any (log-linear) approximations of the utility function, and so we leave the entire structure of the model intact. To ease estimation, we consider a stationary representation of Epstein-Zin preferences, by scaling the life-time utility function by the consumption level. We show how our proposed estimator by can be applied in an asset pricing context. We obtain a sequence of the continuation utility by repeated substitution into the Nadaraya-Watson estimator and, subsequently, we compute the sequence of the stochastic discount factor. To obtain estimates for the parameters of the stochastic discount factor, we minimize a criterion function that is a weighted average of the squared conditional pricing errors. Thereby, this criterion function implies dynamic restrictions on the pricing errors, such that it is zero if and only if the conditional pricing errors are zero at all dates at the estimated parameter values.

We evaluate the performance of our estimator by estimating the parameters of the stochastic discount factor and inspecting the pricing errors. For simplicity, we consider consumption growth to be the single conditioning variable, i.e. the only regressor in the Nadaraya-Watson regressions. We use monthly U.S. data from 1952m2 to 2019m1 for aggregate consumption growth, the six size- and value-based Fama–French bivariate sorted test portfolios of U.S. publicly traded equities, and the return on 3-month
T-Bills. We find plausible values for the parameter of risk aversion and the parameter of the elasticity intertemporal substitution, while the estimate for the parameter of the subjective discount rate is somewhat low. However, we also find suggestive evidence that the parameter for the elasticity intertemporal substitution is weakly identified. The large sample properties are beyond the scope of this paper, but we evaluate the small sample properties of our estimator by simulating a standard long-run risks model. We find that the estimate for the parameter of risk aversion has a small upward bias, that the estimate for the subjective discount rate has a small downward bias, and that the estimate for the elasticity of intertemporal substitution is imprecise.

Our paper is outlined as follows. In section 2 we discuss the relevant literature. We embed this review in a general model specification of recursive preferences, and restrain the attention to the Epstein-Zin functional when necessary. In section 3 we introduce the estimator, explain how we exploit the flexibility of non-parametric regressions, and we explain the criterion function we minimize to estimate the parameters. In sections 4 and 5 we, respectively, describe our dataset and the results. We discuss the obtained estimates of the parameters, as well as the ability of the SDF to price the test assets and the risk-free rate. We will inspect both the joint conditional pricing errors as well as the individual conditional pricing errors. We conduct a Monte Carlo experiment to characterize the finite sample properties of our estimator, and we discuss this experiment in section 6. We conclude in section 7.

2 A general model specification of Recursive preferences with Epstein-Zin preferences

2.1 The dynamic stochastic environment

We consider the following dynamic stochastic environment. We assume time is discrete with dates \( t \in \mathbb{N} \). An event \( z_t \) is drawn from a finite set of events \( \mathcal{Z} \), following an initial event \( z_0 \). The \( t \)-period history of events is denoted by \( z'_t := (z_0, z_1, \ldots, z_t) \) and the set of possible histories is denotes by \( \mathcal{Z}' \subset \mathcal{Z} \). A representative agent has preferences over the payoffs \( c(z'_t) \) for each of the possible histories of \( z'_t \). The general set of preferences are defined by \( V \left( \{c(z'_t)\} \right) \). We consider the Epstein-Zin preferences, defined below.

**Definition 1.** Epstein-Zin preferences are defined on the “Chew-Dekel class” of stationary recursive preferences:

\[
V_t = F(C_t, R_t(V_{t+1})),
\]

where \( V_t \) is the utility starting at \( t \) for history \( z'_t \), \( V_{t+1} \) is the utility for histories \( z'^{t+1} \), \( F : \mathbb{R}^2_+ \mapsto \mathbb{R}_+ \) is the “time aggregator”, and \( R_t \) is a risk aggregator where the subscript \( t \) indicates that the function is based on information available at time \( t \), i.e. the conditional probabilities \( p(z_{t+1} | z'_t) \).\(^3\) Epstein-Zin preferences

\(^3\)We use the notation \( \mathbb{R}_+ \) to denote the positive domain of the set of real numbers, which includes zero, and we use the notation \( \mathbb{R}_{++} \) to denote the strictly positive domain, which excludes zero.
are defined on (1) by:

\[ F(x, y) = \left( (1 - \beta)x^{1 - \frac{1}{\psi}} + \beta y^{1 - \frac{1}{\psi}} \right)^{\frac{1}{1 - \psi}}, \]  

(2a)

\[ R_t(x_{t+1}) = G^{-1} \left( E_t \left[ G(x_{t+1}) \right] \right), \]  

(2b)

\[ G(x) = \frac{x^{1 - \gamma}}{1 - \gamma}, \]  

(2c)

with \( \beta \in (0, 1) \) denoting subjective discount factor, \( E_t \) denoting the expectation conditional on information available at time \( t \), \( \gamma > 0 \) denoting the parameter of risk aversion, and \( \psi > 0 \) denoting the inverse elasticity of intertemporal substitution. Equivalently, Epstein-Zin preferences correspond to:

\[ V_t = \left( (1 - \beta)C_t^{1 - \frac{1}{\psi}} + \beta E_t \left[ V_{t+1}^{1 - \gamma} \right]^{\frac{1}{1 - \gamma}} \right)^{\frac{1}{1 - \psi}}. \]  

(3)

The risk aggregator, thus, corresponds to \( R_t(V_{t+1}) = E_t \left[ V_{t+1}^{1 - \gamma} \right]^{\frac{1}{1 - \gamma}} \), which is a certainty equivalence measure of \( V_t \). Because \( G \) is a concave function, the risk aggregator is lower when \( V_{t+1} \) is more volatile.

In various macroeconomic settings, (1) is the core of the Bellman equation that characterizes the optimal consumption path conditional on the information set \( \mathcal{I}_t \) that is characterized by a model that represents the macroeconomy.\(^4\)

Let us now consider the asset pricing implications of Epstein-Zin preferences. Throughout this paper, boldface symbols denote vectors, matrices, and vector valued functions. Let us assume that the representative investor faces the following intertemporal budget constraint:

\[ C_t + q_t' P_t \leq Y_t + q_{t-1}' \bar{P}_t, \]  

(4)

where the \( l \)-vector \( q_t \) has typical elements \( q_{jt}, j = 1, ..., l \), that denote the quantity of asset \( j \) held at the end of date \( t \), the \( l \)-vector \( P_t \) has typical elements \( P_{jt} \) that denote the prices of asset \( j \) at time \( t \), \( \bar{P}_t \) has typical elements \( \bar{P}_{jt} \) that denote the payoff of asset \( j \) at the start of time \( t \), and \( Y_t \) is an (potentially stochastic) endowment received at time \( t \). We can consider the dynamic programming context in which (1) is the core of the Bellman equation that is subject to (4). The set of control sequences is given by \( \{ (C_t, q_t') \} \). Then, by invoking the Benveniste-Scheinkman Theorem, first-order optimality of the consumption stream implies following conditions (Epstein and Zin, 1989):

\[ \forall t : t_t = E_t \left[ M_{t+1} R_{t+1} \right], \]  

(5)

where \( t_t \) is an \( l \)-vector of ones, \( R_{t+1} := (\text{diag}(P_t))^{-1} \bar{P}_{t+1} \) and the one-day stochastic discount factor (SDF) is given by

\[ M_{t+1} = \frac{F_2(C_t, V_{t+1}) F_1(C_{t+1}, V_{t+2})}{F_1(C_t, V_{t+1})}, \]  

(6)

\(^4\)This setting nests various utility functions that are defined on the set of preferences defined by (1) by means of assuming functional forms of \( F \) and \( R \) (e.g. see Backus et al., 2004).
where $F_i(\cdot, \cdot)$ denotes the first-order derivative w.r.t. the $i$th argument of $F$, $i = 1, 2$. Under the assumption of Epstein-Zin preferences, the SDF is given by

$$M_{t,t+1} := \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\frac{1}{\gamma}} \left( \frac{V_{t+1}}{E_t \left[ V_{t+1}^{1-\gamma} \right]^{1-\gamma}} \right)^{-(\gamma - \frac{1}{\gamma})}. \tag{7}$$

### 2.2 The implications of recursive utility models in constructing estimators

Under various regularity assumptions and restrictions, the first-order condition (5) naturally maps as a moment condition into the Generalized Method of Moments (GMM) framework proposed by Hansen (1982). For instance, imposing $\gamma = 1/\psi$ onto (7) yields the CRRA stochastic discount factor,

$$M_{t,t+1} = \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\gamma}, \tag{8}$$

which is estimated using GMM by Hansen and Singleton (1982). Obviously, this restriction is not one we would like to impose. Then, the problem with recursive preferences is that some of these regularity conditions are not satisfied, since, by construction, the cross-sectional restrictions are multi-period through the recursivity of $V$. For instance, for the Epstein-Zin preferences, the term

$$\mu_{t,t+1} := \left( \frac{V_{t+1}}{E_t \left[ V_{t+1}^{1-\gamma} \right]^{1-\gamma}} \right)^{-(\gamma - \frac{1}{\gamma})}, \tag{9}$$

is not observed as a matter of consequence. A solution to this would be to specify a functional form for $V_{t+1}$. Epstein and Zin (1991), among others, circumvent this issue by taking first order conditions w.r.t. to both $C_t$ and $q_t$. Doing so, yields the following expression for the SDF:

$$M_{t,t+1} = \beta^\theta \left( \frac{C_{t+1}}{C_t} \right)^{-\frac{\theta}{\psi}} (R^{A}_{t+1})^{\theta-1}, \tag{10}$$

where $\theta := (1-\gamma)/(1-1/\psi)$ and $R^{A}_{t+1}$ denotes the return on the aggregate consumption claim (that is, the optimal portfolio implied optimal values of $q_t$). Indeed, by using a proxy for $R^{A}_{t+1}$, the Hansen (1982) GMM framework can be applied, e.g. see Epstein and Zin (1991).

Even though Epstein-Zin preferences are nowadays the most popular specification in macro-finance modeling, and dynamic asset pricing in particular, they do not remain free of controversy. These controversies arise because of various identification issues. First, perhaps the biggest debate is about the parameter values. For instance, when $\gamma > 1/\psi$, the preferences imply “early resolution of risk” (Kreps and Porteus, 1978; Bansal et al., 2010). Though intuitively appealing, the data do not necessarily imply

5By equating (7) and (10) and rewriting, we can find an expression for $\mu_{t,t+1}$ in terms of $R^{A}_{t+1}$,

$$\mu_{t,t+1} = (\beta R^{A}_{t+1})^{\theta-1} \left( \frac{C_{t+1}}{C_t} \right)^{\frac{1-\theta}{\psi}}, \tag{11}$$

which we can interpret as a functional form for $\mu_{t,t+1}$. 

5
this parameter restriction because the obtained estimates for both $\gamma$ and $\psi$ are often imprecise (e.g. see Bansal and Yaron, 2004; Chen et al., 2013). Moreover, some point out that i) the implied willingness to pay for the early resolution of risk is implausibly high under conventional calibrations (Epstein et al., 2014), and ii) common calibrations suggest that much of the variation in asset prices is driven by fears that the economy will do bad in the far future rather than prompt economic fluctuations, which they consider to be implausible (e.g. see Cochrane, 2017). Next to that, risk premia, their volatilities, and patterns of return predictability are extremely sensitive to certain parameter values of the assumed endowment process, particularly when one makes assumptions about the persistence of shocks to consumption growth (Hansen et al., 2008; Pohl et al., 2018a,b).

These controversies are difficult to resolve due to econometric challenges and, consequently, theoretical models are often calibrated with little to no econometric guidance of the parameter values. For instance, one challenge pertains to the use of proxies for the return on the wealth portfolio, $R^A_{t+1}$. The derivation of (10) rests on a complete markets assumption. Therefore, $R^A_{t+1}$ does not only include returns on stocks and bonds, but also on non-perishable consumption goods – such as cars, stamp collections, and art – and on human capital. The model might be rejected because the proxy is not sufficiently broad so that it does not capture the wealth portfolio. This notion is an extension of Roll’s (1977) critique, who applies this logic to the Capital Asset Pricing Model. He argues that a model specification test that involves a proxy on the wealth portfolio constitutes a joint test on the model specification and the adequacy of the proxy. This is not a light issue; Lettau and Ludvigson (2001) estimate that about two-thirds of total wealth consists of human capital.

A second challenge arises from the fact that consumption growth is not that volatile and (close to) independent and identically distributed. Because consumption growth is not that volatile, the parameter $\psi$ is difficult to estimate precisely. For instance, $1.001^{2}$ and $1.001^{2/3}$ are not so wildly different in values. Notwithstanding, $\psi = 0.5$ or $\psi = 1.5$ have wildly different asset pricing implications and economic substance (e.g. see Bansal and Yaron, 2004; Bansal et al., 2010). Next to that, the measure of certainty equivalence,

$$E_t \left[ V_{t+1}^{1-\gamma} \right]^{\frac{1}{1-\gamma}},$$  \tag{12}$$
is rendered meaningless when consumption growth is i.i.d., but in the data consumption growth is (close to) i.i.d. and, therefore, the power of classical test is low (Chen et al., 2013).\footnote{We will later show in Proposition 1 that this term is indeed rendered meaning less when consumption growth is i.i.d., because then $V_{t+1}/C_{t+1}$ is constant. This, in turn, implies that the term $\mu_{t+1}$ equals 1 for all $t$, and Epstein-Zin preferences collapse to a CRRA specification.}

A third challenge is that the parameters are oftentimes estimated in a Hansen (1982) GMM environment, and that the instruments used are oftentimes weak. Some have tried to resolve these econometric challenges by proposing novel confidence sets. For instance, Stock and Wright (2000) suggest confidence sets that are immune to weak instruments and Yogo (2004) derives valid confidence intervals of a linearized specification of the SDF. However, Manresa et al. (2017) show that such strategies do not perform well in small samples. Moreover, linearization works well for preferences close to log utility, i.e. $\gamma \approx 1$ and $\psi \approx 1$. If risk aversion indeed is close to 10, a value commonly used in calibrations, linearization techniques are not at all appropriate. Kleibergen and Zhan (2018) generalize the approach of Stock and Wright (2000) as to include many risk factors in both linear and non-linear specifications, and they allow
for joint tests on the pricing errors. However, the confidence sets they propose remain to be rather wide.

Various studies consider the estimation of long-run risk models, such as Bansal et al. (2007, 2010, 2016); Constantinides and Ghosh (2011); Grammig and Küchlin (2018); Meddahi and Tinang (2016). Essentially, in a long-run risk model the evolution continuation value, $V_{t+1}$, is characterized by the assumed processes for consumption growth and dividend growth. For instance, Meddahi and Tinang (2016) propose a procedure for estimating the Bansal and Yaron (2004) long-run risk model that allows for the separation between the investor’s optimal decision frequency and the frequency of the data. This technique is robust to weak identification. However, their estimator can be applied to the Bansal and Yaron (2004) model only and has “difficulty to match many moments”. Grammig and Küchlin (2018) propose a two-step indirect inference approach to estimate the long-run risk model that separates the estimation of the model’s macro-economic dynamics and the preference parameters. Our study also relates to these studies in that we do need to specify a macroeconomic model. We propose a strategy in which there is no need to restrict processes for consumption growth nor dividend growth. As such, we can estimate and test a broad set of models because we do not require a complete and explicit representation of the stochastic equilibrium.

Our study contributes to the literature that in that we simultaneously circumvent the need to specify a macroeconomic model as well as Roll’s critique by avoiding the use of a proxy for the return on the wealth portfolio. Our work is, therefore, closest to Chen et al. (2013), who also employ a semi-parametric technique to estimate (7). They proceed in two steps. First, they identify state variables over which the continuation value is identified by making mild restrictions on the dynamic behavior of consumption growth. Specifically, they first assume that consumption growth is a linear function of a hidden univariate first-order Markov process, $x_t$, and they assume the joint distribution of consumption growth and $x_t$ to be normal and independent and identically distributed. Second, they show that the ratio $V_t/C_t$ can be fully characterized by function $F$ of lagged consumption growth, $C_t/C_{t-1}$, and $V_{t-1}/C_{t-1}$. Third, they approximate the unknown function $F$ by a bivariate sieve function $\hat{F}$. Finally, they generate a sequence of $V_t/C_t$ using $\hat{F}$ and minimize Hansen’s (1982) GMM criterion to estimate the parameters $(\beta, \gamma, \psi)$. We use a (non-parametric) kernel density estimator to estimate the continuation value. This approach differs in four subtle but important ways. First, our assumption on the consumption growth process is even more relaxed. We do not have to make any assumptions about the distribution of consumption growth and our proposed estimator accommodates weak serial dependence in consumption growth. Second, we include the complete distribution of consumption growth, rather than the point estimate of the conditional expectation of the continuation utility based on one lag of consumption growth and the continuation utility. Our approach, however, does not inform us which regressors (i.e. state variables) should be used in the non-parametric regressions. Third, our approach does not involve any instruments, and thereby we mitigate issues pertaining to weak instruments and instrument validity. Fourth, we use the complete structure of the model because we do not have to use (log-linear) approximations to estimate the conditional expectation nested in the risk aggregator.

3 A semi-parametric approach to estimating models with recursive preferences

To alleviate the issues presented in the previous section, we use a semi-parametric approach. Specifically, we leave the parametric specification of the SDF (7) intact, and model the evolution of in the continuation
utility using a Nadaraya-Watson kernel regression.

We consider a discrete-time setting, again indexed by \( t \), with the \((d_1 \times d_2)\)-dimensional stochastic process \((X_t, Y_t)\). Let \( \Omega \) denote the set of sample points in the underlying probability space that we consider in our estimation. The filtration, \( F_t \), generated by the process \( \{X_n\}_{n=0}^t \) represents the flow of available information to the investor at time \( t \).

**Assumption 1.** Under the physical (real) probability measure \( \mathbb{P} : \mathcal{F} \mapsto [0, 1] \), the process \( \{X_t\} \) is stationary and ergodic. Specifically, \( \{X_t\} \) is time-homogeneous and Markov of order 1 in \( \mathcal{X} \subset \mathbb{R}^{d_1} \), with transition density \( f(x_t | x_{t-1}) \).

We consider a sample of size \( T \), which is defined by a finite segment of one realization of the process \( \{X_t, Y_t\} \), i.e. \( \{X_t(\omega_0), Y_t(\omega_0) : 1 \leq t \leq T\} \) for some \( \omega_0 \in \Omega \). For instance, in the recursive utility framework with a representative investor, the sample includes the growth of aggregate consumption (in \( X_t \)) and the returns on a set of test-portfolios (in \( Y_t \)). We consider a set of cum-dividend gross \( t + 1 \) returns collected in the \( q \)-vector \( R_{t+1} \).

Next, we consider the compact space \( \Theta \subset \mathbb{R}^k \). We wish to estimate the \( k \)-vector \( \theta \in \Theta \) using data observed from a sample over the interval \( \mathcal{T} = \{1, \ldots, T\} \). Let us assume that there exists a one-day SDF, \( M_{t,t+1} \) that prices the universe of assets.

**Assumption 2.** Markets are absent of arbitrage, and so there exists a stochastic scalar process \( \{M_{t,t+1}\} \) that:

1. \( M_{t,t+1} > 0 \);
2. is measurable with respect to the information \( F_{t+1} \);
3. \( M_{t,t+1} \in \mathcal{M} \), where \( \mathcal{M} \) is the set of admissible square integrable SDFs, and;
4. the sequence \( \{M_{t,t+1}\} \) satisfies the following no arbitrage restrictions:

   \[
   E \left[ M_{t,t+1} R_{t+1} - i_q | X_t = x \right] = 0_q \tag{13}
   
   \]

for any \( x \in \mathcal{X} \), where \( i_q \) is a \( q \)-vector of ones and \( 0_q \) is a \( q \)-vector of zeros.

Additionally, we assume the following.

**Assumption 3.** The one-day SDF \( M_{t,t+1} \) is a \( F_{t+1} \) measurable function \( m : \mathcal{X} \times \Theta \mapsto \mathbb{R}^{++} \) of the state-variables at \( t + 1 \) and the parameters \( \theta \), i.e. \( M_{t,t+1} = m(X_{t+1}; \theta) \).

Next, we consider the procedure that alleviates the problems sketched in section 2.2. First, we define \( v_t := V_t/C_t \) and \( G_t := C_t/C_{t-1} \), and subsequently we reformulate (1) to a function of stationary variables by rescaling \( V_t \) with \( C_t \):

\[
\begin{align*}
v_t &= \left( 1 - \beta + \beta E \left[ (v_{t+1})^{1-\gamma} \left( G_{t+1} \right)^{1-\gamma} \bigg| X_t = x \right] \right)^{\frac{1}{1-\gamma}} X_t^{\frac{1}{1-\gamma}} \bigg| \left( \frac{1}{1-\gamma} \right) + \frac{1}{1-\gamma} \bigg]. \tag{14}
\end{align*}
\]
Assumption 1 permits us to condition on only one lag of \( X \). The stochastic discount factor (7) can be written in terms of \( v \) as follows:

\[
M_{t,t+1} = \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\frac{1}{\psi}} \sqrt{\frac{V_{t+1} C_{t+1} C_i}{C_{t+1} C_i}} \frac{1}{\psi} \left( \frac{V_{t+1} C_{t+1} C_i}{C_{t+1} C_i} \right)^{1-\gamma} \left( \frac{X_t = x}{1-\gamma} \right) \]

\[
= \beta \left( G_{t+1} \right)^{-\frac{1}{\psi}} \frac{v_{t+1} G_{t+1}}{E \left[ G_{t+1} \right]^{1-\gamma} X_t = x} \left( \frac{X_t = x}{1-\gamma} \right),
\]

Now, \( \{v_t\} \) is stationary sequence, and a function of the stationary sequences \( \{G_t\} \) and \( \{v_t\} \), which eases the estimation. Next, we consider the following proposition.

**Proposition 1.** When \( \Delta \ln(C_{t+1}) \overset{i.i.d.}{\sim} \mathcal{N} (\mu, \sigma^2) \), and \( \lim_{T \to \infty} v_T = \bar{v} \), then \( \bar{v} \) equals:

\[
\bar{v} = \frac{1 - \beta}{1 - \bar{x}} \xi := \beta \exp \left\{ \left( 1 - \frac{1}{\psi} \right) \mu + \frac{(1 - \gamma)(1 - 1/\psi \sigma^2)}{2} \right\}.
\]

**Proof.** See Appendix A.

Thus, when log-consumption growth is independent and identically distributed, the continuation utility is rendered obsolete. Naturally, we will not assume that this is the case. But it turns out that this proposition is useful in the estimation, that is discussed next.

### 3.1 A nonparametric estimator for the continuation utility

To estimate the parameters \( \theta := (\beta, \gamma, \psi') \), we introduce the following strategy. We compute \( \forall \theta \in \Theta \subset (0, 1) \times \mathbb{R}^2_+ \), iterating over \( i = 0, 1, \ldots, N \):

\[
v(x; \theta, i) = \begin{cases} 
\left( 1 - \beta + \beta \varphi(x; i - 1, \theta) \right)^{1-\psi} \left( \frac{1}{1-\sigma} \right)^{1-\psi}, & i > 0 \\
\left( \frac{1-\beta}{1-\sigma} \right)^{1-\psi}, & i = 0,
\end{cases}
\]

where

\[
\varphi(x; \theta, i - 1) := E \left[ v(G_{t+1}; \theta, i - 1) \right]^{1-\gamma} \left( G_{t+1} \right)^{1-\gamma} \left( X_t = x \right).
\]

We estimate the values for \( \varphi \), backwards from \( i = 0 \) to \( i = N \) using a Nadaraya-Watson kernel regression (Nadaraya, 1964; Watson, 1964). We confine our analyses to the case where \( X_t = G_t \) for the sake of simplicity. It is possible to include more conditioning information nonetheless. However, the
number of conditioning variables one can include is limited because product kernels suffer from the curse of dimensionality. We motivate the chosen value for $v(x; \theta, 0)$ by Proposition 1. Indeed, the assumption that consumption growth is i.i.d. is not an assumption that we will carry throughout the analyses. Instead, we conjecture that consumption growth is sufficiently close to i.i.d., such that has a mean not far from $\bar{v}$. That is to say, we consider the assumed terminal value, which equals (16), to be a sufficiently adequate approximation of the mean of $\{v(X_t; \theta, N)\}_{t=1}^{T-1}$. Essentially, we compute

$$\hat{\phi}(G; \theta, i - 1) := \begin{cases} \int_C v(G_{t+1}; \theta, i - 1)^{1-\gamma}(G_{t+1})^{1-\gamma} \hat{f}(G_{t+1} | G)dG_{t+1}, & i > 1, \\ \left(\frac{1-\beta}{1-\gamma}\right) \int_C (G_{t+1})^{1-\gamma} \hat{f}(G_{t+1} | G)dG_{t+1}, & i = 1, \end{cases}$$

(19)

working iteratively backwards from $i = 1$ to $i = N$, where at each $i$ we compute $\hat{\phi}$ for all $G_t$ using an estimated transition density $\hat{f}$, and we define

$$\hat{\xi} := \beta \exp \left\{ (1 - \frac{1}{\psi}) \hat{\mu} + \frac{(1-\gamma)(1-1/\psi)}{2} \hat{\sigma}^2 \right\},$$

(20)

where $\hat{\mu}$ and $\hat{\sigma}^2$ are, respectively, the sample mean and variance of the observed log-consumption growth series. However, for the minimization of computational costs, we directly consider a kernel regression.

**Assumption 4.** We assume the following about the univariate kernel function $K : \mathbb{R} \mapsto \mathcal{K}$, $\mathcal{K} \subseteq \mathbb{R}_+$:

$$\int K(z)dz = 1,$$

(21a)

$$K(z) = K(-z) \Rightarrow \int zK(z)dz = 0,$$

(21b)

$$\int z^2K(z)dz = \tau_2 > 0.$$

(21c)

The Nadaraya-Watson kernel regression associated with (18) is given by:

$$\hat{\phi}(G; \theta, i - 1, h_T) := \sum_{j=1}^{T-1} K \left( \frac{G_j - G}{h_T} \right) \hat{v}(G_{j+1}; \theta, i - 1)^{1-\gamma} (G_{j+1})^{1-\gamma} \left/ \sum_{k=1}^{T-1} K \left( \frac{G_k - G}{h_T} \right) \right.,$$

$$= \sum_{j=1}^{T-1} \hat{w}(G, G_j; h_T) \hat{v}(G_{j+1}; \theta, i - 1)^{1-\gamma} (G_{j+1})^{1-\gamma},$$

(22)

where $K$ denotes the kernel function, $h_T$ denotes the kernel bandwidth, the (kernel) weights are defined as

$$\hat{w}(G, G_j; h_T) := K \left( \frac{G_j - G}{h_T} \right) \left/ \sum_{k=1}^{T-1} K \left( \frac{G_k - G}{h_T} \right) \right.,$$

(23)

\text{We indeed find that when we take } \hat{\eta}(x; \theta, 0) = 0, \text{ we converge to the same estimate for sequence } \{v_t\} \text{ as when we take } \hat{\eta}(x; \theta, 0) = \bar{v}, \text{ where } \bar{v} \text{ is defined by Proposition 1. However, because } \bar{v} \text{ is close to the mean of } \hat{\eta}(x; \theta, N), \text{ we need far less number of iterations over } i \text{ before convergence when we start at } \bar{v} \text{ instead of } 0. \text{ This difference in number of iterations is particularly large for low values of } \gamma, \text{ because a low value of } \gamma, \text{ say 2, implies a low curvature of the risk aggregator, and so } \bar{v} \text{ is relatively closer to the mean of } \hat{\eta}(G; \theta, N) \text{ compared to large values of } \gamma, \text{ say 20.}
and we have defined $\hat{v}$ as follows:

$$
\hat{v}(G_{j+1}; \theta, i, h_T) :=
\begin{cases}
(1 - \beta + \beta \hat{\phi}(G_{j+1}; \theta, i - 1, h_T))^{1 - \frac{1}{\hat{\phi}}} , & i > 0 \\
(\frac{1 - \beta}{1 - \hat{\phi}})^{1 - \frac{1}{\hat{\phi}}} , & i = 0.
\end{cases}
$$

(24)

This estimator is based on the prehension that (14) defines a contraction mapping w.r.t. $v_t$, and so we should obtain a consistent estimate of the sequence $\{v_t\}$ by repeated substitution of $v(\cdot; i)$ for $i = 1, 2, \ldots$. When one takes a long horizon $N$, $\{v(G_i; \theta, N)\}$ should converge, with which we mean that:

$$
||v(x; \theta, i) - v(x; \theta, i - 1)|| \leq \epsilon(\cdot),
$$

(25)

where $\epsilon$ is a strictly positive real-valued function that characterizes the consistency of the kernel estimator. For norms on various other forms of contraction mappings, such convergence properties are documented (e.g. see Anatolyev, 1999; Bosq, 1996; Györfi et al., 2013). Characterization of the convergence properties of an iterative estimate is oftentimes analytically intractable and must, therefore, be considered separately for the problem and algorithm at hand. Characterization of the convergence properties of our estimator is beyond the scope of this paper, and we leave them for future work.

### 3.2 Evaluating the moment conditions

To estimate the parameter vector $\theta$, we minimize the quadratic conditional pricing errors using some (conditional) weighting matrix. We consider three different weighting matrix to evaluate the robustness of our results: the identity matrix, the inverse of the conditional covariance matrix of the gross returns, and the inverse of the conditional second moment of the gross returns matrix. The identity matrix weighs the pricing errors equally, while the inverses of the conditional covariance matrix of gross returns and the conditional matrix of second moments of gross returns provide an economically interesting weighting matrix. First, we describe the criterion function. Second, we discuss the empirical implementation of the various matrices. Third, we discuss issues pertaining to the inversion of the near-singular covariance matrix and the second moment matrix, and discuss how we mitigate these issues.

#### 3.2.1 The Criterion Function

We evaluate the ability of the SDF to price the assets conditionally. Therefore, we evaluate conditional pricing errors using dynamic restrictions. That is, we estimate the parameter vector $\theta$ by a criterion function of the form:

$$
\delta = \min_{\theta \in \Theta} E \left[ e(X_i; \theta)^T \Omega(X_i) e(X_i; \theta) \right]^{1/2}.
$$

(26)

where

$$
e(X_i; \theta) := E \left[ h(X_{i+1}, R_{i+1}; \theta) | X_i \right].
$$

(27)
is the $q$ dimensional vector of conditional pricing errors defined by the vector valued function

$$h(X_{t+1}, R_{t+1}; \theta) := m(X_{t+1}; \theta) R_{t+1} - t_q.$$  

We compute the conditional expectation of the pricing errors, $e(X_t; \theta)$, using the conditioning information $X_t$. The positive definite (p.d.) matrix $\Omega(X_t)$ is the weighting matrix that weighs the conditional pricing errors, and its value can depend on the conditioning information $X_t$.

We would like to highlight that this criterion imposes *dynamic* restrictions on the pricing errors because, we square the conditional pricing errors (with the weighting matrix in between) before taking the expectation. That is to say, $\delta$ is zero if and only if the pricing errors equal zero at each time $t$. In contrast, the classical GMM criterion associated with the Sargan-Hansen statistic, for instance, is zero when the pricing errors are on average zero – and thus does not imply dynamic restrictions on the pricing errors.

We use various specifications for the weighting matrices $\{\Omega(X_t)\}$. For instance, the identity matrix, $I$, weighs all pricing errors equally both within the cross-section as well as across time. We also consider the inverse of the matrix of conditional second moments of the gross returns. Hansen and Jagannathan (1997) consider a GMM framework, and motivate the inverse of the (unconditional) second moment matrix of gross returns as a weighting matrix as a distance measure between a model for the SDF and the space of true SDFs. We follow the approach of Gagliardini and Ronchetti (2019), in that we compute the *dynamic* restrictions of the pricing errors using the sequence of conditional second moment of gross return matrices; this is the Conditional Hansen-Jagannathan (HJ) Distance. Specifically, we define the set $\mathcal{M}$ of admissible SDFs for the chosen test assets’ gross returns:

$$\mathcal{M} := \{ M_{t,t+1} \in L^2(F_{t+1}) : E \left[ M_{t,t+1} R_{t+1} - t_q \mid X_t = x \right] = 0_q \},$$  

where $L^2(F_{t+1})$ denotes the linear space of random variables with finite second moment and measurable w.r.t. $F_{t+1}$. Gagliardini and Ronchetti (2019) define the Conditional HJ Distance, $\delta$, on a set of *dynamic* pricing restrictions as follows:

$$\delta := \min_{\theta \in \Theta} \min_{M_{t,t+1} \in \mathcal{M}} E \left[ \left\{ M_{t,t+1} - m(X_{t+1}; \theta) \right\}^2 \right]^{1/2}.$$  

(30)

The following proposition follows from solving the inner constrained optimization problem on the RHS of (30).

**Proposition 2.** The Conditional HJ Distance is equal to:

$$\delta = \min_{\theta \in \Theta} E \left[ e(X_t; \theta)^\prime \Omega(X_t) e(X_t; \theta) \right]^{1/2},$$  

(31)

where

$$e(X_t; \theta) := E \left[ h(X_{t+1}, R_{t+1}; \theta) \mid X_t \right],$$  

(32)

is the $q$ dimensional vector of conditional pricing errors defined by the vector valued function

$$h(X_{t+1}, R_{t+1}; \theta) := m(X_{t+1}; \theta) R_{t+1} - t_q.$$  

(33)
and $\Omega(X_t) := E \left[ R_{t+1} R'_{t+1} | X_t \right]^{-1}$ is the inverse of the matrix of conditional second moments of $R_{t+1}$.

**Proof.** See Appendix A.1 in Gagliardini and Ronchetti (2019). □

Thus, the Conditional HJ Distance is a special case nested in (26) in which $\Omega(X_t) = E \left[ R_{t+1} R'_{t+1} | X_t \right]^{-1}$. The Conditional HJ Distance metric does not involve a set of instruments, unlike the (unconditional) Hansen-Jagannathan (1997) metric:

$$\delta_Z := \min_{\theta \in \Theta} \left( E \left[ Z(X_t) h(X_{t+1}, R_{t+1}; \theta) \right]' \Omega_Z(X_t) E \left[ Z(X_t) h(X_{t+1}, R_{t+1}; \theta) \right] \right)^{1/2}, \quad (34)$$

where $Z(X_t)$ is a $p \times q$ matrix of instruments, $p \geq k$, and

$$\Omega_Z(X_t) := E \left[ Z(X_t) R_{t+1} R'_{t+1} Z(X_t)' \right]^{-1}. \quad (35)$$

We also consider the inverse of the conditional covariance matrix of returns, $\nabla( R_{t+1} | X_t)^{-1}$. The second moment matrix is difficult to invert because it is near singular. We use the conditional covariance matrix as a robustness check on the impact of numerical inversion inaccuracies.

### 3.2.2 Nonparametric estimation of the weighting matrices

We follow the same strategy as Gagliardini and Ronchetti (2019), and so we estimate the Conditional HJ Distance by taking the sample average of (31) and by estimating the conditional pricing errors (32) by a kernel regression. Specifically, we estimate the pricing error vector $e(X_t; \theta)$ by a Nadaraya-Watson regression:

$$\hat{e}_T(X_t; \theta) := \sum_{i=1}^{T-1} w(X_t, X_i; h_T) \hat{h} \left( X_{t+1}, R_{t+1}; \theta \right). \quad (36)$$

where we compute

$$\forall t \in T_0, \theta \in \Theta : h(X_{t+1}; \theta, N) := \hat{m}(X_{t+1}; \theta) R_{t+1} - 1_q, \quad (37)$$

with $\Theta$ a discrete three-dimensional grid of admissible parameter values for $(\beta, \gamma, \psi)$, and

$$\hat{m}(X_{t+1}; \theta) := \beta \left( G_{t+1} \right)^{-\frac{1}{\psi}} \left( \frac{\hat{v}(G_{t+1}; \theta, N) G_{t+1}}{\hat{q}(G_{t+1}; \theta, N - 1)^{1/\tau}} \right)^{\frac{\psi - 1}{\psi}}, \quad (38)$$

which is the empirical counterpart of (15).

The corresponding sample equivalent of the criterion function, $\delta_T$, is defined as:

$$\hat{\delta}_T^2 := \min_{\theta \in \Theta} D_T(\theta) = D_T(\hat{\theta}_T), \quad D_T(\theta) := \frac{1}{T} \sum_{t=1}^{T} \| (X_t) e_T(X_t; \theta)' \hat{\Omega}_T(X_t) e_T(X_t; \theta), \quad (39)$$

where (39) defines $\hat{\theta}_T$ and $\hat{\Omega}_T(X_t)$ is the sample equivalent of $\Omega(X_t)$. The indicator variable $1(X_t)$ equals one when $X_t \in X' \subset X$, where $X'$ is a given compact subset of the state variables’ support, and $1(X_t)$
is zero when \( X_t \notin \mathcal{X}_\delta \). The indicator variable is a trimming factor that controls boundary effects in the kernel regression (e.g. see Tripathi and Kitamura, 2003; Su and White, 2014). Gagliardini and Ronchetti (2019) show that, since the criterion function \( D \) involves \( 1(X_t) \), the estimator \( \hat{\delta}_T \) is not a consistent estimator of \( \delta \), but of

\[
\delta_* := \min_{\theta \in \Theta} E \left[ \hat{\upsilon}(X_t)e(X_t; \theta')\Omega(X_t)e(X_t; \theta) \right]^{1/2}
\]

(40) instead. Let \( \theta_* \) be the minimizer of (40), and we assume \( \theta_* \) to be unique in \( \Theta \). Naturally, \( \delta_* \leq \delta \), but \( \theta_* \) coincides with true parameter vector \( \theta_0 \) if the SDF is correctly specified (Gagliardini and Ronchetti, 2019).

Next, we consider the sample equivalents of the various weighting matrix, \( \Omega_T(X_t) \). When we use an identity matrix as a weighting matrix, we weigh all pricing errors equally, both within the cross-section as well as across time. We obtain the sample equivalent of the Conditional HJ Distance weighting matrix by:

\[
\hat{\Omega}_T(X_t) = \left( \sum_{i=1}^{T-1} w(X_t, X_i; h_T) R_{i+1}^T R_{i+1}^\prime \right)^{-1}
\]

(41) which is the kernel regression estimator of the matrix \( E \left[ R_{i+1}^T R_{i+1}^\prime | X_t \right]^{-1} \). To obtain the conditional variance matrix of the gross returns, \( \Psi \), we use the law of total covariance to decompose the variance matrix:

\[
\Psi \left( R_{t+1} | X_t \right) = E \left[ R_{t+1} R_{t+1}^\prime | X_t \right] - E \left[ R_{t+1} | X_t \right] E \left[ R_{t+1}^\prime | X_t \right].
\]

(42)

Then, we note that the first element on the RHS is the \( q \times q \) conditional second moment matrix, which we can estimate by (42), and that the \( q \)-vector \( E \left[ R_{t+1} | X_t \right] \) and its transpose can be estimated by a Nadaraya-Watson regression. Therefore, the non-parametric estimator for \( \Psi(R_{t+1} | X_t)^{-1} \) is given by:

\[
\hat{\Psi}_T \left( R_{t+1} | X_t \right)^{-1} = \left( \sum_{i=1}^{T-1} w(X_t, X_i; h_T) R_{i+1}^T \right) \left( \sum_{i=1}^{T-1} w(X_t, X_i; h_T) R_{i+1}^\prime \right)^{-1}.
\]

(43)

### 3.3 Kernel function and bandwidth selection

Some remarks on the Nadaraya-Watson regression are in order. For reasons that will become clear below, the subscript \( T \) of the bandwidth \( h_T \) used in the estimation of the conditional expectation indicates that the bandwidth depends on the size of the sample, and converges to zero when \( T \to \infty \). Additionally, we postulate that \( h_T T \to \infty \) when \( T \to \infty \), which ensures that the size of “local sample” increases when the sample size increases. Third, we choose the Gaussian kernel,

\[
K(z) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z^2}{2} \right),
\]

(44)

because it is simple to implement and the performance of the kernel estimator depends substantially on the chosen bandwidth, and not so much on the chosen kernel \( K \) (e.g. see Li and Racine, 2007).
Let us now turn to the selection of the bandwidth $h_T$. For the sake of exposition, I drop function arguments and subscripts where no confusion can arise. The Nadaraya-Watson regression (22) estimates the conditional expectation:

$$E\left[\varphi \mid G\right] = \int \varphi \frac{f(G, \varphi)}{f(G)} d\varphi,$$

where $f(x, y)$ is a joint density of $x$ and $y$. Equivalently, we can say we consider a non-parametric regression equation

$$\varphi(G) = g(G) + u,$$

where $u$ is an error term and the estimator of $g$ is denoted by $\hat{g}$ and corresponds to (22). For exposition sake, let us consider a general environment in which we have a set of observations $\{Y_t, X_t\}_{t=1}^T$, where $Y_t$ is a dependent variable and $X_t$ is an independent variable. Indeed, $Y$ can be the outcome of a function. We can evaluate the performance of the estimator by evaluating the mean squared error (MSE):

$$MSE(x, h_T) := E\left[\left(\hat{g}(x, h_T) - g(x)\right)^2\right] = \text{var}(\hat{g}(x, h_T)) + \text{bias}\left\{\hat{g}(x, h_T), g(x)\right\}^2.$$

In the setting where we only consider one regressor, the approximate MSE is given by:

$$MSE(x, h_T) \approx \frac{\text{var}(y|x)}{f(x)T h} \int K^2(z) dz + h^2 \left(\frac{1}{2}g''(x) + \frac{g'(x)f'(x)}{f(x)}\right) \tau_2,$$

where the first term on the RHS is the approximate variance and the second term on the RHS is the approximate bias (Pagan and Ullah, 1999). The MSE is a local measure, since it depends locally on $x$. Therefore, we consider the integrated mean squared error (IMSE), that is defined by

$$\text{IMSE}(h_T) := \int MSE(x, h_T) dx.$$

The optimal bandwidth minimizes the IMSE (e.g. see Li and Racine, 2007). The (approximate) variance is decreasing in $T$ and $h_T$, while the (approximate) bias is increasing in $h_T$, and so the optimal bandwidth depends on both the bandwidth as well as the sample size. That is to say, the optimal bandwidth balances variance and bias given a sample size $T$ – hence, the notation $h_T$, rather than just $h$.

Various methods and routines to estimate the optimal bandwidth exist (e.g. see Bosq, 1996; Li and Racine, 2007). We discuss three popular methods here. For a univariate Nadaraya-Watson regression, one way to pick $h$ is

$$h_{opt}^{rt} := C \times \hat{s}_x T^{-1/5}, \quad C > 0,$$

where $\hat{s}_x$ is the estimated standard deviation of $x$ and $C$ is a constant.\(^8\) A rule of thumb is to use $C = 1.059$ or $C = 1$ when one uses a Gaussian kernel (e.g. see Bosq, 1996; Li and Racine, 2007).\(^9\) A similar

---

\(^8\)In a multivariate settings, such rules do not perform well (Li and Racine, 2007, p. 66–67)

\(^9\)Other ways pick $C$ are labeled plug-in methods, which are not discussed here, but can be found in various textbooks on nonparametric econometrics.
bandwidth is suggested by Silverman (1986):

\[
h_{opt}^s := 0.9 \times \min \left\{ \sigma_x, \frac{r_q(x)}{1.34} \right\} T^{-1/5},
\]

where \( r_q(x) \) is the interquartile range of \( x \). These optimal bandwidths are based on the IMSE in which the density of \( x \) is estimated.

Two popular data driven methods outperform this rule-of-thumb, but are difficult to apply in our context. One data driven method is Least Squares Cross-Validation (CV), in which the optimal bandwidth is obtained

\[
h_{opt}^{CV} := \arg \min_h \frac{1}{T} \sum_{i=1}^{T} (Y_i - \hat{g}_{-i}(X_i, h))^2,
\]

where \( \hat{g}_{-i} \) denotes the estimator of \( g \) in which the observation \( X_i \) has been left out. Another popular method is introduced by Hurvich et al. (1998), which is based on an improved Akaike Information Criterion (AIC):

\[
\begin{align*}
\text{AIC}_c(X, y; h) &:= \ln (\hat{\sigma}^2(X, y; h)) + \frac{1 + \text{tr}(W(X; h)) + 2}{T}, \\
\hat{\sigma}^2(X, y; h) &:= \frac{Y'(I - W(X; h))'(I - W(X; h))Y}{T},
\end{align*}
\]

where \( X := (X_1, \ldots, X_T)' \), \( Y := (Y_1, \ldots, Y_T)' \), and \( W : T \times T \) has typical elements

\[
W_{ij} = \frac{K((X_i - X_j)/h)}{\sum_{k=1}^{T} K((X_i - X_k)/h)},
\]

i.e. \( W \) is the matrix with the kernel weights. The AIC optimal bandwidth is given by:

\[
h_{opt}^{AIC} := \arg \min_h \text{AIC}_c(x, y; h).
\]

Hurvich et al. (1998) and Li and Racine (2004) show that the CV method and AIC method have equivalent asymptotic performance but that the AIC method performs better in small samples.

At first glance, it would be natural to apply the CV or AIC method in our setting because it is pivotal to obtain a good prediction of the conditional expectation of the continuation utility. However, we iteratively update \( \phi \) over the horizon \( i = 0, \ldots, N \) because it is unobserved. Perhaps we could update our CV or AIC bandwidth within the estimation process, but our prediction of \( \nu \) is based on the value of the bandwidth as well as on the value of the parameters we estimate. Because we have not characterized the convergence properties of our estimator, we adopt the Silverman (1986) bandwidth (51). Additionally, we also use the Silverman (1986) bandwidth to compute \( \hat{\Phi}_T(X_i) \) and \( \hat{\epsilon}_T(X_i; \theta) \) to minimize computational costs.
4 Data

We estimate the model’s parameters using U.S. data from 1959m2 to 2019m1. The representative agent’s consumption growth is proxied by the growth rate of the sum of the series Personal Consumption Expenditures: Nondurables and Personal Consumption Expenditures: Services, which both are obtained via the Federal Reserve Economic Data (FRED) database maintained by the Federal Reserve of St. Louis. To account for the assumed representative agent framework, we scale the consumption level by the size of the U.S. population using the series Population that is also taken from the FRED database.

To evaluate the ability of the stochastic discount factor to price returns, we consider a proxy for the risk-free rate and sets of cum-dividend gross returns on test portfolios. We proxy the gross risk-free rate with the gross return on 3-Month T-Bills, which we have obtained from FRED. The test portfolios are obtained from Kenneth French’s website.\footnote{\url{http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html}} We consider six value weighted Frama-French portfolios, which are two-way sorted along the size and book-to-market factors. The constructed portfolios are classified along the book-to-market dimension as small (G), neutral (N), and value (V), and along the market capitalization dimension as small (S) and big (B). Accordingly, we label the gross time $t$ returns by $SG_t$, $SN_t$, $SV_t$, $BG_t$, $BN_t$, $BV_t$. We denote the time $t$ gross return on the 3-month T-Bill by $R^f_t$.

To adjust for inflation we use the Personal Consumption Expenditures Chain-type Price Index (PCEPI) to obtain both real consumption growth as well as the set ex-post real returns on the test portfolios. The PCEPI series has also been obtained from FRED. Additionally, we augment some of the figures we will present with shaded areas that represent the National Bureau of Economic Research (NBER) recession periods. These have been created with the NBER recession indicators, that we have also obtained from FRED.

The descriptive statistics are exhibited in Table 1.

\begin{table}[h]
\centering
\begin{tabular}{c|ccccc}
\hline
 & Mean & SD & Minimum & Maximum & Skewness & Excess Kurtosis \\
\hline
$G_t$ & 1.001 & 0.008 & 0.961 & 1.032 & -0.256 & 2.567 \\
$SG_t$ & 1.006 & 0.066 & 0.674 & 1.267 & -0.367 & 1.763 \\
$SN_t$ & 1.009 & 0.053 & 0.718 & 1.260 & -0.463 & 2.689 \\
$SV_t$ & 1.011 & 0.055 & 0.721 & 1.298 & -0.353 & 3.185 \\
$BG_t$ & 1.006 & 0.045 & 0.766 & 1.205 & -0.355 & 1.727 \\
$BN_t$ & 1.006 & 0.042 & 0.795 & 1.162 & -0.367 & 2.102 \\
$BV_t$ & 1.008 & 0.048 & 0.782 & 1.206 & -0.389 & 2.291 \\
$R^f_t$ & 1.006 & 0.004 & 0.989 & 1.024 & 0.735 & 1.174 \\
\hline
\end{tabular}
\caption{Descriptive statistics for data from 1959m2 to 2019m1 ($T = 720$)}
\end{table}

SD is the sample standard deviation and the excess Kurtosis is the Kurtosis in excess relative to the normal distribution, i.e. the Kurtosis minus 3.

5 Results

Our program can be set two routines, to minimize the sample criterion function. The first one is a brute force grid search, in which we make $m$ combinations of the vector $\theta$ that are within $\Theta$ and compute which grid point returns the minimizer of (39). This method is rather computationally expensive, especially...
when the grid is fine, but ensures a solution that is close to the global optimum within $\Theta$. Another method is to use a minimization routine, such as Gauss-Newton, that finds $\hat{\theta}_T$ by minimizing (39). It is, however, difficult to determine a priori whether the function we attempt to minimize is unimodal. Minimization routines, however, rely on an initial guess around which it searches for an optimum. Thus, the minimization routine might return a local minimizer. To ensure that the minimization routines searches around the global minimum, we employ a Genetic Search algorithm to find an approximate solution to the global minimum. Then, we pass the solution obtained by the Genetic Search algorithm as a first guess to the minimizer routine. The details of the Genetic Search algorithm are explained in Appendix B.

A well documented issue is the near singularity of both the covariance matrix of returns as well as the second moment matrix of returns (e.g. see Cochrane, 2009, p. 213–217). The reason for this near singularity is that risky assets exhibit high positive correlation, in particular at short intervals such as monthly returns. Because numerical inversions are imprecise for near-singular matrices, we perform three different inversions to evaluate the robustness of our point estimates to the inversion. First, we use a regular numerical inversion routine, and use the imprecise estimates for the sequence $\{\hat{\Omega}_T(X_t)\}$. Second, we use a Moore-Penrose pseudo-inversion to obtain the sequence $\{\hat{\Omega}_T(X_t)\}$. The third method is a pseudo-inversion based on shrinking the concentrations ratios of the uninverted matrices. The concentration ratio of any p.d. matrix $Z$ is defined as:

$$\text{CR}(Z) := \sqrt{\frac{\lambda_{\max}(Z)}{\lambda_{\min}(Z)}},$$

where $\lambda_{\max}(Z)$ is the largest eigenvalue of $Z$ and $\lambda_{\min}(Z)$ is the smallest eigenvalue of $Z$. To obtain $\{\hat{\Omega}_T(X_t)\}^T_{t=1}$, we inflate the diagonals of the matrices $\{\hat{\Omega}_T(X_t)^{-1}\}^T_{t=1} - i.e. the matrices before inversion – by a constant $\hat{c} > 0$ such that the concentration ratios $\{\text{CR} (\hat{\Omega}_T(X_t))^{-1} + \hat{c}I\}$ are close to 15.11 That is, $\hat{c} = \arg\min \hat{c} \sum_{t=1}^{T} [15 - \text{CR} (\hat{\Omega}_T(X_t))^{-1} + \hat{c}I]^2$.

We would like to highlight that this method does not affect the correlation structure of the gross returns on the test assets.

The point estimates using the various weighting matrices and inversions are exhibited in Table 2. We do not document standard errors, because we do not have large sample properties of our estimator and a bootstrap is virtually impossible because our predicted sequence $\{\hat{v}(G_t; \hat{\theta}_T, N)\}^T_{t=1}$ is a function of the parameter vector $\theta$. To evaluate the small sample properties of our estimates, we perform a Monte Carlo experiment, and we discuss this in section 6. The estimates based on the identity matrix as a weighting matrix are exhibited in column 1. The identity matrix weights all conditional pricing errors equally both within the cross-section as well as in the time series. The estimate for the subjective discount rate, $\hat{\delta}$, equals 0.952, and that is rather low considering that we use monthly data. The estimate for the parameter of risk aversion, $\hat{\gamma}$, equals 1.229, which is relatively low as well, yet reasonable from an economic perspective. This low value suggests that consumers are somewhat risk averse, but rather willing to change consumption between “states of nature”. The estimate for the elasticity of intertemporal substitution, $\hat{\psi}$, is relatively low as well and, notably, it is below 1. This low value suggests that the

\[\text{Perhaps a more accurate way would be to find a unique } \hat{c} \text{ for each } \hat{\Omega}_T(X_t). \text{ However, that is rather computationally intensive and, as it has turned out, a uniform } \hat{c} \text{ yields concentration ratios close to 15.} \]
Table 2: Estimates of $\hat{\theta}_T =: (\hat{\beta}, \hat{\gamma}, \hat{\psi})'$ using various weighting matrices, $\hat{\Omega}_T(X_t)$, and numerical inversion methods to obtain $\hat{\Omega}_T(X_t)$.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}$</td>
<td>0.952</td>
<td>0.966</td>
<td>0.945</td>
<td>0.967</td>
<td>0.950</td>
<td>0.999</td>
<td>0.995</td>
</tr>
<tr>
<td>$\hat{\gamma}$</td>
<td>1.229</td>
<td>15.515</td>
<td>10.231</td>
<td>15.461</td>
<td>16.056</td>
<td>4.185</td>
<td>4.345</td>
</tr>
<tr>
<td>$\hat{\psi}$</td>
<td>0.812</td>
<td>7.365</td>
<td>4.096</td>
<td>9.109</td>
<td>4.386</td>
<td>3.053</td>
<td>2.985</td>
</tr>
<tr>
<td>$\delta_T$</td>
<td>0.017</td>
<td>0.264</td>
<td>0.297</td>
<td>0.264</td>
<td>0.298</td>
<td>0.026</td>
<td>0.035</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Omega(X_t)$</th>
<th>I</th>
<th>CVar</th>
<th>CSM</th>
<th>CVar</th>
<th>CSM</th>
<th>CVar</th>
<th>CSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inversion</td>
<td>Normal</td>
<td>Normal</td>
<td>MP</td>
<td>MP</td>
<td>R</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>$\hat{\epsilon}$</td>
<td>0.3416</td>
<td>0.0317</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We denote the weighting matrix $\Psi((R_{t+1}\mid X_t)^{-1}$ by CVar (Conditional Variance), and the weighting matrix $\mathbb{E}[R_{t+1}R_{t+1}'\mid X_t]^{-1}$ by CSM (Conditional Second Moment). The Moore-Penrose inversion is denoted by MP, and R denotes that the matrix inverse is computed by regularization, which implies that the matrix is inverted by inflating the diagonal elements of the uninverted matrices by $\bar{\epsilon}$ to shrink the concentration ratio. We have used the inter-ventile range as a trimming factor $X_\star$, which implies that $\mathbb{I}(X_t) = 1$ when $X_t$ is within the 0.95 innerquantile range of $X$ and that $\mathbb{I}(X_t) = 0$ otherwise.

The interest rate is rather sensitive to shocks in consumption growth because consumers are not that willing to substitute consumption over time.

Column 2 of Table 2 exhibits the point estimates based on the inverse of conditional covariance matrix of gross returns, $\hat{\Psi}(R_{t+1}\mid X_t)^{-1}$, that is inverted using a standard numerical inversion routine. We find that the estimate for the subjective discount rate, 0.966, is still relatively low because we consider monthly data. The point estimate for the parameter of risk aversion, 15.515, is much higher than in column 1, but it remains to be reasonable. The estimate for the intertemporal elasticity of substitution is also much higher, and perhaps a bit too high given that it implies that the interest rate is extremely insensitive to changes in consumption growth. Because the matrices $\hat{\Psi}(R_{t+1}\mid X_t)$ are near singular, the computed inverse is rather imprecise. Columns 4 and 6 are a robustness check on these results by means of another inversion. Column 4 exhibits the results obtained when $\hat{\Psi}(R_{t+1}\mid X_t)$ is inverted using a Moore-Penrose pseudo-inversion. We find that the results are similar to the estimates in column 2. Indeed, the point estimate for the elasticity of intertemporal substitution, 9.109, is substantially higher. However, the criterion function is flat around $\psi$, which suggests that it is weakly identified and so the estimates do probably not differ significantly. Column 6 exhibits the estimates that are obtained when the matrices $\hat{\Psi}(R_{t+1}\mid X_t)$ are inverted by inflating their diagonal elements by $\bar{\epsilon} = 0.3416$ to shrink the concentration ratios of the matrices $\hat{\Psi}(R_{t+1}\mid X_t)$. Then, we find that the estimate for the subjective discount rate equals 0.999, which is a value we would expect given that we use monthly data. The estimate for the parameter of risk aversion equals 4.185, which is a reasonable value. The point estimate for the elasticity of intertemporal substitution equals 3.053 is somewhat high, but remains to be plausible from an economic perspective.

Column 3 of Table 2 exhibits the point estimates based on the inverse of conditional second moment matrix of gross returns, $\hat{\mathbb{E}}[R_{t+1}R_{t+1}'\mid X_t]^{-1}$, that is inverted using a standard numerical inversion routine. This weighting matrix corresponds to the Conditional Hansen-Jagannathan Distance. We find that the
estimate for the subjective discount rate, 0.945, is also rather low because we consider monthly data. The
point estimate for the parameter of risk aversion, 10.231, is substantially higher than in column 1, but is
considered to be reasonable. The estimate for the intertemporal elasticity of substitution equals 4.096
which is relatively high, but reasonable. Because the conditional second moment matrices of gross returns
is near singular, the computed inverse is rather imprecise. Columns 5 and 7 are a robustness check on
these results by means of using other inversion techniques to obtain the weighting matrices. Column
5 exhibits the results obtained when the matrices \( \hat{E}[R_{t+1} R'_{t+1} | X_t] \) are inverted using a Moore-Penrose
pseudo-inversion. We find that the results are similar to the estimates in column 3, except that the estimate
for the parameter of risk aversion in column 5, 16.056, is substantially higher than in column 3. Column 6
exhibits the estimates that are obtained when the matrices \( \hat{E}[R_{t+1} R'_{t+1} | X_t] \) are inverted by inflating their
diagonal elements by \( \hat{\sigma} = 0.0317 \) to shrink their concentration. Then, we find that the estimate for the
subjective discount rate equals 0.995, which is a value we would expect given that we use monthly data.
The estimate for the parameter of risk aversion equals 4.345, which is a reasonable value. Additionally, the
criterion function is highly curved around \( \gamma \), which suggests that it is well identified. The point estimate
for the elasticity of intertemporal substitution equals 2.985, which perhaps somewhat high, but remains to
be plausible from an economic perspective. However, the criterion function is rather flat around \( \hat{\psi} \), which
suggests that it is weakly identified.

5.1 Dynamic behavior of the stochastic discount factor and the pricing errors
Let us now turn to the evaluation of the dynamics of the SDF and the pricing errors. For the sake of
conciseness, we focus on the dynamic pricing errors we have obtained using the conditional second
moment matrix of gross returns as the weighting the matrix in the criterion function, where we have
inverted the matrix by regularization. We consider these results because this weighting matrix corresponds
to the Conditional Hansen-Jagannathan Distance criterion function, which is economically interesting,
and because we think that inversion by regularization is more precise than a normal inversion or a Moore-
Penrose pseudo-inversion. Figure 1 exhibits the evolution of consumption growth, \( G_{t+1} \), and the shaded
areas represent the NBER recession periods. Because we control the boundary effects with the indicator
function \( 1(G_t) \), which we set to 0 for value outside the inner-ventile range, we do not condition on, for
example, the recession in 2008. Additionally, we see that there is no strong volatility clustering and
 persistence, which suggests that consumption growth is close to i.i.d. indeed.

![Figure 1: Consumption growth \( G_t \), over time \( t \).](image)

Figure 2 exhibits the evolution of the one-day SDF, \( \hat{M}_{t,t+1} := m(G_{t+1}; \hat{\theta}_t) \), which we decompose in
the scaled consumption growth term $G_{t+1}^{-1/\hat{\nu}}$, and the state variable

$$\hat{\mu}_{t,t+1} := \left( \frac{\hat{\nu}(G_{t+1}; \theta, N) G_{t+1}}{\hat{\varphi}(G_{t+1}; \theta, N - 1)} \right)^{-\frac{1}{\hat{\nu}}}.$$  (58)

We can see in the top panel of Figure 2 that the SDF exhibits some heteroskedasticity, and is relatively volatile around recessions. We can derive from the middle and bottom panels of this figure that the lion’s share of the variance and the heteroskedasticity can be attributed to variation in $\hat{\nu}_{t,t+1}$, not in $G_{t+1}^{-1/\hat{\nu}}$. This might explain why we find a point estimate for $\hat{\nu}$ that is sufficiently low. For instance, the SDF corresponding to a time-separable life-time utility function with an instantaneous CRRA utility function, $M_{t,t+1} = \beta G_{t+1}^{-\gamma}$, has difficulty to match the equity premium; that is, one as to assume a very high level of risk aversion, say 90, to generate sufficient variation in the SDF to match the equity premium (Cochrane, 2017; Mehra and Prescott, 1985). Because the component $\hat{\nu}_{t,t+1}$ is rather volatile, we do not have to inflate the variation in consumption growth by raising it to a large power; that is, a reasonable value for $\hat{\nu}$ of 2.985 suffices.

5.2 Inspection of the pricing errors

We also consider the time-series of the statistic $1(X_t)\hat{\nu}(X_t; \hat{\theta}_T)\hat{\nu}(X_t; \hat{\theta}_T)$, which we exhibit in Figure 3. In panel a of Figure 3, we exhibit the pricing errors where we take $1(X_t) = 1$ for all $t$. We see that the joint pricing errors are very large around both the burst of the dotcom bubble as well as in the 2008-2009 Great Financial Crisis. At this point, however, $1(X_t) = 0$ in the computation of the criterion function because at these points $X_t \notin \mathcal{X}_\delta$. Therefore, we exhibit the series $1(X_t)\hat{\nu}(X_t; \hat{\theta}_T)\hat{\nu}(X_t; \hat{\theta}_T)$ with $1(X_t) = 0$ when $X_t \notin \mathcal{X}_\delta$ in panel b of Figure 3. We can see in panel b that the joint pricing errors are rather small. Next to that, the model does not perform distinctively well or poor in some relative to other periods. Indeed, we have used this indicator function to control boundary effects of our estimator, and it might be that the kernel estimates of $\nu$, $\mathbf{Q}$, and $e$ are rather poor in these extreme periods where consumption growth is relatively low. However, given that the model performs poorly in the periods where $1(X_t) = 0$, which we can see by comparing the two panels, and the the value of $\hat{\nu}(X_t; \hat{\theta}_T)\hat{\nu}(X_t; \hat{\theta}_T)$ is quite large in these periods, our estimates might be biased. For instance, we might need to generate more volatility in the SDF to match the price fluctuations in these periods by raising $\gamma$.

Let us now turn to the individual pricing errors, which we have exhibited in Figure 4. A typical challenge for cross-sectional models in asset pricing is to price various test assets implied by the Fama-French two-way sort well (cf. Cochrane, 2011, 2017). To ease comparison of the pricing errors, the y-axes of the panels in Figure 4 are aligned. We can see in the panels of Figure 4 that the pricing errors of some assets are not substantially more volatile than the pricing errors of others. Again, we see that the pricing errors around the dotcom bubble and, in particular, the 2008 Great Financial Crisis are large. In Table 3 we report the time-series averages of the individual conditional pricing errors, their standard deviation, and their correlations. The time series averages of the pricing errors are rather close to each other, and so are their volatilities. However, the average pricing errors are all slightly positive, even though the criterion function is lowest at $\hat{\theta}_T$. We can attribute this to the inversion of the weighting matrix, which

\[12\]Mehra and Prescott (1985) suggest a lower bound of 4 and an upper bound of 10 for $\gamma$. 

21
FIGURE 2: Dynamics of the one-day stochastic discount factor process $\hat{M}_{t,t+1} := m(X_{t+1}; \hat{\theta}_t)$.

The dynamics of the discount factor are based on $\hat{\theta}_t$ that is computed by minimizing the criterion function with $\hat{\theta}_t(X_t) = \left( \sum_{i=1}^{T-1} w(X_t, X_i; h_T)R_{t+i}R'_{t+i} \right)^{-1}$, where the inverse is obtained by regularization. We used the 90 innerquantile for the trimming factor $I(X_t)$.
The joint conditional pricing errors are computed with the estimate $\hat{\theta}_T$ that is computed by minimizing the criterion function with $\Omega_T(X_t) = \left( \sum_{t=1}^{T-1} \psi(X_t, h_t R_{i,t+1}, R'_{i,t+1}) \right)^{-1}$, where the inverse is obtained by regularization. We used the 90 innerquantile for the trimming factor $1(X_t)$. 

**Figure 3**: The joint conditional squared pricing errors over time $t$. 

(a) Time series of $\tilde{e}_T(X_t; \hat{\theta}_T)^T \Omega_T(X_t) \tilde{e}_T(X_t; \hat{\theta}_T)$. 

(b) Time series of $1(X_t) \tilde{e}_T(X_t; \hat{\theta}_T)^T \Omega_T(X_t) \tilde{e}_T(X_t; \hat{\theta}_T)$. 


The scalar $\hat{e}_{T,j}(X_t; \hat{\theta}_T)$ is the $j$’th typical element of the vector $\hat{e}_T(X_t; \hat{\theta}_T)$, $j = 1, \ldots, q$. The joint pricing errors are computed with the estimate $\hat{\theta}_T$ that is computed by minimizing the criterion function with $\hat{\theta}_T(X_t) = \left( \sum_{i=1}^{T-1} a(X_t, X_i; h_T) R_{it} R'_{i+1} \right)^{-1}$, where the inverse is obtained by regularization.

We used the 90 innerquantile for the trimming factor $l(X_t)$. 

**Figure 4**: Pricing errors $\hat{e}_{T,j}(X_t; \hat{\theta}_T)$ of the various test portfolios and the 3-month T-Bill, $j$, over time $t$. 

---

24
might be imprecise. Because the conditional second moment matrix is near singular, its inverse has large negative off-diagonal elements. Nevertheless, that the pricing errors have about the same time-series average and volatility implies that indeed the SDF does not have trouble pricing one particular asset. The pricing errors exhibit a strong positive correlation, which implies that it is difficult to price them jointly at particular points in times.

**Table 3**: Sample time series properties of the conditional pricing errors

<table>
<thead>
<tr>
<th></th>
<th>$\hat{e}_{SG}$</th>
<th>$\hat{e}_{SN}$</th>
<th>$\hat{e}_{SV}$</th>
<th>$\hat{e}_{BG}$</th>
<th>$\hat{e}_{BN}$</th>
<th>$\hat{e}_{BV}$</th>
<th>$\hat{e}_{TBILL3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.006</td>
<td>0.009</td>
<td>0.010</td>
<td>0.006</td>
<td>0.008</td>
<td>0.008</td>
<td>0.006</td>
</tr>
<tr>
<td>SD</td>
<td>0.012</td>
<td>0.009</td>
<td>0.010</td>
<td>0.006</td>
<td>0.007</td>
<td>0.009</td>
<td>0.005</td>
</tr>
<tr>
<td>Correlations</td>
<td>$\hat{e}_{SG}$</td>
<td>0.941</td>
<td>0.863</td>
<td>0.709</td>
<td>0.492</td>
<td>0.494</td>
<td>0.350</td>
</tr>
<tr>
<td></td>
<td>$\hat{e}_{SN}$</td>
<td>0.974</td>
<td>0.712</td>
<td>0.630</td>
<td>0.692</td>
<td>0.343</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{e}_{SV}$</td>
<td>0.710</td>
<td>0.698</td>
<td>0.794</td>
<td>0.393</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{e}_{BG}$</td>
<td>0.871</td>
<td>0.761</td>
<td>0.587</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{e}_{BN}$</td>
<td>0.926</td>
<td>0.602</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{e}_{BV}$</td>
<td></td>
<td>0.480</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The constructed portfolios are classified along the Fama-French book-to-market dimension as small (G), neutral (N), and value (V), and along the market capitalization dimension as small (S) and big (B). Accordingly, we label the pricing error with the subscripts $SG$, $SN$, $SV$, $BG$, $BN$, $BV$. The label $TBILL3$ corresponds to the 3-month T-Bill.

### 6 Monte Carlo experiment

In this section we evaluate the finite sample performance of our proposed estimator by means of a Monte Carlo experiment. We use the Bansal and Yaron (2004) long-run risks model as the data generating process (DGP), which we describe in detail in section 6.1. We describe our evaluation and numerical implementation in section 6.2 and we discuss the results of the Monte Carlo experiment in section 6.3.

#### 6.1 The data generating process

We assume the Bansal and Yaron (2004) economic environment that incorporates fluctuating economic uncertainty. We give a brief summary of the model here to harmonize the notation with our paper and for the sake of completeness. Let small caps denote the natural logarithms of capitals, e.g. $g_{t+1} := \ln(G_{t+1})$. We consider the SDF representation (10), which is mathematically equivalent to (7), and we restate it here for the sake of convenience:

$$\forall j : \mathbb{E}_t \left[ \exp \left\{ \theta \ln(\beta) - \frac{\theta}{\psi} g_{t+1} + (\theta - 1)r^{sd}_{t+1} + r_{j,t+1} \right\} \right] = 1.$$  

(59)
To connect asset prices to macroeconomic uncertainty, Bansal and Yaron (2004) model log-consumption
growth, \( g_t \), and log-dividend growth, \( g_{d,t} \), as follows:

\[
g_{t+1} = \mu + x_t + \sigma_t \eta_{t+1}, \tag{60a}
\]

\[
g_{d,t+1} = \mu_d + \phi x_t + \varphi_d \sigma_t u_{t+1}, \tag{60b}
\]

\[
x_{t+1} = \rho x_t + \varphi_e \sigma_t \epsilon_{t+1}, \tag{60c}
\]

\[
\sigma^2_{t+1} = \tilde{\sigma}^2 + v_1 (\sigma^2_t - \tilde{\sigma}^2) + \sigma_w \psi_{t+1}, \tag{60d}
\]

where

\[
(\eta_{t+1}, u_{t+1}, \epsilon_{t+1}, \psi_{t+1})' \sim \mathcal{N}(0, I). \tag{60e}
\]

The component \( x_t \) in the growth terms is persistent and predictable, particularly when \( \rho \in (0, 1) \) is close to one and \( \sigma_t \) is “small”. Additionally, note that consumption growth and dividend growth are i.i.d. when \( \sigma_w = \varphi_e = 0 \). The parameters \( \phi > 1 \) and \( \varphi_d > 1 \) can be used to calibrate the volatility of dividends and its correlation with dividend growth.

Let us now characterize the dynamics of the return processes. Bansal and Yaron (2004) show that using a standard Campbell-Shiller (1988) decomposition of returns approximations, we have that:

\[
r_{at+1} = \kappa_0 + \kappa_1 z_{t+1} - z_t + g_{t+1}, \tag{61}
\]

where \( z_t := \ln(P_t) - \ln(C_t) \) and \( \kappa_1 = \exp(\bar{z})/ (1 + \exp(\bar{z})) \), with \( \bar{z} \) being the linearization point and \( P_t \) the price of the consumption claim. The return on the wealth portfolio, i.e. the consumption claim, can typically not be observed. Nevertheless, one can also consider the observed return on an aggregate dividend claim, e.g. a stock market index, \( R_{t+1}^m \). The Campbell-Shiller decomposition of that return is as follows:

\[
r_{at+1} = \kappa_{0,m} + \kappa_{1,m} z_{m,t+1} - z_{m,t} + g_{d,t+1}, \quad \kappa_{1,m} = \exp(\bar{z}_m)/ (1 + \exp(\bar{z}_m)) \tag{62}
\]

where \( g_{d,t} \) is the growth rate of the aggregate dividend, \( D_t \), and \( z_{m,t} := \ln(P^m_t) - \ln(D_t) \), with \( P^m_t \) the price of the aggregate dividend claim. Using the method of undetermined coefficients, (61), and the no-arbitrage restriction (59), we find that \( z_{t+1} \) obeys:

\[
z_{t+1} = A_0 + A_1 x_{t+1} + A_2 \sigma^2_{t+1}, \tag{63}
\]

where

\[
A_0 = \frac{1}{1 - \kappa_1} \left[ \ln(\psi) + \kappa_0 + \left( \frac{1 - \frac{1}{\psi}}{\psi} \right) \mu + \kappa_1 A_2 (1 - v_1) \tilde{\sigma}^2 + \frac{\theta}{2} (\kappa_1 A_2 \sigma_w)^2 \right], \tag{64}
\]

\[
A_1 = \frac{1 - \frac{1}{\psi}}{1 - \kappa_1 \rho}, \tag{65}
\]

\[
A_2 = \frac{1}{2} \left[ \left( \frac{\theta - \psi}{\psi} \right)^2 + (\theta A_1 \kappa_1 \varphi_e)^2 \right] / \theta (1 - \kappa_1 v_1), \tag{66}
\]

26
and using (62) we find that $z_{m,t+1}$ obeys:

$$z_{m,t+1} = A_{0,m} + A_{1,m}x_{t+1} + A_{2,m}^2\sigma^2_{t+1},$$

where

$$A_{0,m} = \frac{\left(\theta \ln(\beta) + (\theta - 1)\kappa_0 + \kappa_{0,m} + (\theta - 1)(\kappa_1 - 1)A_0 + \left(\theta - 1 - \frac{\theta}{\psi}\right)\mu + \mu_d\right)}{1 - \kappa_{1,m}^2},$$

$$A_{1,m} = \frac{\phi - \frac{1}{\psi}}{1 - \kappa_{1,m}^2},$$

$$A_{2,m} = \frac{(1 - \theta)(1 - \kappa_1\nu_1)A_2 + \frac{1}{2}H}{(1 - \kappa_{1,m}\nu_1)},$$

with

$$H := \lambda_{m,\eta}^2 + (\beta_{m,e} - \lambda_{m,e})^2 + \phi_d^2,$$

$$\beta_{m,e} := \kappa_{1,m}A_{1,m}\varphi_e,$$

$$\beta_{m,w} := \kappa_{1,m}A_{2,m},$$

$$\lambda_{m,e} := (1 - \theta)\kappa_1\varphi_e,$$

$$\lambda_{m,w} := (1 - \theta)A_2\kappa_1,$$

$$\lambda_{m,\eta} := -\gamma.$$  

Bansal and Yaron (2004) show how to obtain $A_1$, $A_2$, $A_{1,m}$, and $A_{2,m}$. We have obtained $A_0$ by substituting $r_{t+1}^d = r_{i,t+1} = \kappa_0 + (\kappa_{1,m} - 1)A_0 + A_1(\kappa_1x_{t+1} - x_t) + A_2(\kappa_1\sigma_{t+1}^2 - \sigma_t^2) + g_{t+1}$ into (59). Similarly, we have obtained $A_{0,m}$ by substituting $r_{i,t+1} = \kappa_{0,m} + (\kappa_{1,m} - 1)A_{0,m} + A_{1,m}(\kappa_{1,m}x_{t+1} - x_t) + A_{2,m}(\kappa_{1,m}\sigma_{t+1}^2 - \sigma_t^2) + g_{d,t+1}$ into (59). To obtain the risk-free rate, $r_{t+1}^f$, we substitute $r_{t+1} = r_{j,t+1}$ into (59) and rewrite to obtain:

$$r_{t+1}^f = -\ln(\beta) + \frac{1}{\psi}E_t\left[g_{t+1}\right] + \frac{1 - \theta}{\theta}E_t\left[r_{t+1} - r_{t+1}^f\right] - \frac{1}{2\theta} \text{var}_t\left(\theta\psi g_{t+1} + (1 - \theta)r_{t+1}^d\right).$$

Then, we note that

$$E_t\left[g_{t+1}\right] = \mu + x_t,$$

$$E_t\left[r_{t+1} - r_{t+1}^f\right] = -\lambda_{m,\eta}\sigma_t^2 + \lambda_{m,e}\kappa_1\varphi_e\sigma_t^2 + \kappa_1\lambda_{m,w}\sigma_t^2 - \frac{1}{2}\text{var}_t(r_{t+1}^d),$$

$$\text{var}_t\left(r_{t+1}^d\right) = \left[1 - (\kappa_1\varphi_e)^2\right]\sigma_{t+1}^2 + (A_2\kappa_1\sigma_{t+1}^2),$$

$$\text{var}_t\left(\theta\psi g_{t+1} + (1 - \theta)r_{t+1}^d\right) = \left(\lambda_{m,\eta}^2 + \lambda_{m,e}^2\right)\sigma_t^2 + \lambda_{m,w}^2\sigma_t^2.$$

and use these results to simplify simulating the risk-free rate.
6.2 The implementation

We evaluate the finite sample properties of the estimator by simulating $B = 999$ time series with length $T = 1,000$. We use a burn in of 10,000 periods to mitigate the potential influence of initial values. We use the same calibration as Bansal and Yaron (2004) Table IV with $\gamma = 10$, which we summarize in Table 4. The parameter values correspond to a monthly calibration. These parameters imply the values for $\kappa_1$ and $\kappa_{1,m}$ which, in turn, imply the values for $\kappa_0$ and $\kappa_{0,m}$.

We consider the linearization point $\bar{z} = \mathbb{E}[z_{t+1}]$, so that $\bar{z} = A_0 + A_2 \bar{\sigma}^2$, since $\mathbb{E}[x_{t+1}] = 0$. Therefore, we know that $\kappa_1$ satisfies:

$$
\kappa_1 = \frac{\exp \left( A_0 + A_2 \bar{\sigma}^2 \right)}{1 + \exp \left( A_0 + A_2 \bar{\sigma}^2 \right)}.
$$

(82)

Similarly, we find that $\kappa_{1,m}$ satisfies:

$$
\kappa_{1,m} = \frac{\exp \left( A_{0,m} + A_{2,m} \bar{\sigma}_m^2 \right)}{1 + \exp \left( A_{0,m} + A_{2,m} \bar{\sigma}_m^2 \right)}.
$$

(83)

Note that $A_0$ and $A_2$ depend on $\kappa_1$, and that $A_{0,m}$ and $A_{2,m}$ depend on $\kappa_{1,m}$, and so given the values for the parameters in Table 4, we use a non-linear equations solver to find the values for $\kappa_1$ and $\kappa_{1,m}$. We find that the values in Table 4 imply that $\kappa_1 = 0.9981$ and that $\kappa_{1,m} = 0.9959$.

Again, we consider consumption growth as the single conditioning variable; $X_t = G_t$. We do so because we want to determine to what extent consumption growth suffices as a conditioning variable and because we want to adopt a uniform approach throughout the paper, even though the relevant state variables are $x_t$ and $\sigma_t^2$. Additionally, this involves the use of product kernels, because then we would consider more than one regressor. Moreover, we could also obtain estimates for $x_t$ and $\sigma_t^2$ from the real data using. We leave such applications for future work.

6.3 The results

We minimize the criterion function (26) for each of the simulated samples $b = 1, ..., B$, where we use the inverse of the conditional second moment matrix as the weighting matrix, i.e. we obtain the estimates by minimizing the Conditional Hansen-Jagannathan Distance. We use regularization to invert the matrix.
Because the SDF should in principle be able to price all assets, we use the gross return on the wealth portfolio, the gross return on the market portfolio, and the gross return on the risk free asset.

![Distribution of $\hat{\beta}$](image1)

![Distribution of $\hat{\gamma}$](image2)

![Distribution of $\hat{\phi}$](image3)

**Figure 5:** The distribution of the estimated SDF parameters.

The estimates for each of the Monte Carlo samples are obtained by minimizing the criterion function with $\hat{\Omega}_T(X_t) = \left( \sum_{i=1}^{T-\tau} \delta(X_t, X_i; h_i) R_{t+i} R_{t+i}' \right)^{-1}$, where the inverse is obtained by regularization. We used the 90 innerquantile for the trimming factor $1(X_t)$.

Figure 5 exhibits the distribution of the estimated SDF parameters. Panel 5a exhibits the distribution of $\hat{\beta}$. The true parameter value $\beta$ equals 0.998 and is represented by the solid vertical line. The median of the 999 bootstrap estimates of $\beta$ is represented by the dashed vertical line. Ideally, the dashed vertical line is on top of the solid vertical line. We see, however, that there is a strong downward bias in the estimates of $\beta$. This is surprising because the $\beta$ should in principle be well identified because it determines the level of the SDF. The dotted line is the kernel density estimate of the probability density function of $\hat{\beta}$. We see that it has the highest mass around the median of the $\hat{\beta}$, but that the density is not unimodal. However, we can also attribute this to the fact that the inverses of $\hat{\Omega}_T$ are imprecise, which might generate a downward bias in the statistic $\hat{\delta}(X_t; \hat{\beta}_{T}) \hat{\Omega}_T(X_t) \hat{\delta}(X_t; \hat{\beta}_{T})$ (cf. section 5.2).

Panel 5b exhibits the distribution of $\hat{\gamma}$. The true parameter value $\gamma$ equals 10 and is represented by the solid vertical line. The median of the 999 bootstrap estimates of $\beta$ is represented by the dashed vertical line. We see, however, that there is a small upward bias in the estimates of $\gamma$. We can attribute this to the fact that the estimate of a sequence that estimated by repeated substitution can have a small downward bias around periods where it has a low variance (e.g. see Anatolyev, 1999). We could resolve this with
using a local polynomial estimator for the sequence, instead of a Nadaraya-Watson regression, or other techniques (e.g. see Härdle, 1990). To correct for this lower variance $\hat{v}$, we might need a higher value for $\gamma$ to inflate the variance of $\hat{\mu}$ to reduce the pricing errors. The dotted line is the kernel density estimate of the probability density function of $\hat{\gamma}$. We find that a lot of mass is at the median of $\hat{\gamma}$, which suggests that the standard errors are small, whereas in existing studies the standard errors for this parameter are rather large (cf. Chen et al., 2013; Epstein and Zin, 1991; Bansal et al., 2010). However, there is also a small bump around 1.8. This is not a boundary issue of the kernel estimator of the density of $\hat{\gamma}$ because the minimization routine occasionally returns values close to 1.8 indeed.

Panel 5c exhibits the distribution of $\hat{\psi}$. The true parameter value $\psi$ equals 1.5 and is represented by the solid vertical line. The median of the 999 bootstrap estimates of $\beta$ is represented by the dashed vertical line, and we see, that there is an upward bias in the estimate of $\psi$. We would like to highlight that it the criterion function is rather flat around $\psi$. The minimization routine often searches initially around a very wide range of values and, subsequently, searches for quite some time around the estimate it finally returns. Likewise, the estimate of the probability density function, the dotted line, is rather flat. It has two peaks, one around the true value and one around 4.2. Given that the criterion function is so flat around $\psi$, we feel we could attribute the atypical distribution of $\hat{\psi}$ to anomalies in the minimization routine.

7 Conclusion

We propose a novel semi-parametric estimator for non-linear rational expectations model with Epstein-Zin preferences. We leave the parametrization of the Epstein-Zin stochastic discount factor intact, and model the evolution of the continuation utility using a kernel estimator. Hereby, we resolve three identification issues that prevail in the existing literature: we circumvent the need of specifying a macroeconomic model or to proxy the return on the wealth portfolio to model the evolution of the continuation utility, and we leave the utility model intact because we do not have to invoke (log-linear) approximations for the evolution of the continuation utility. We show how one can obtain a sequence for the continuation utility by repeated substitution into a Nadaraya-Watson estimator. Additionally, we show how the obtained sequences for the continuation utility and the risk aggregator can be used compute the sequence for the Epstein-Zin stochastic discount factor. This representation of the stochastic discount factor can, in turn, be used to estimate the parameters of the Epstein-Zin stochastic discount factor by minimizing an appropriate criterion function.

We evaluate the performance of our estimator by estimating the parameters of the stochastic discount factor and inspecting the pricing errors. Using monthly U.S. data from 1952m2 to 2019m1 for aggregate consumption growth and the six size- and value-based Fama–French bivariate sorted test portfolios of U.S. publicly traded equities. We minimize a criterion function that imposes dynamic restrictions on the conditional pricing errors. We weigh the conditional pricing errors using various weighting matrices; the identity matrix, the inverse of the conditional covariance matrix of gross returns, and the conditional second moment matrix of gross returns, which corresponds to minimizing the Conditional Hansen-Jagannathan Distance. Because the covariance matrix and the second moment matrix are near singular, we also consider various inversion routines for and evaluate the impact of the chosen inversion routine. We find plausible values for the parameter of risk aversion, while the estimate for the parameter of the subjective discount rate is somewhat low. Because the criterion function we minimize, is flat around the parameter for the elasticity of intertemporal substitution, we consider the estimate for the
elasticity of intertemporal substitution to be weakly identified. We find that the values of the estimates are somewhat robust to the chosen weighting matrix – though the estimates based on the identity matrix differ substantially from the other estimates – but can be sensitive to the chosen inversion of the weighting matrix. We evaluate the small sample properties of our estimator by simulating a standard long-run risks model. We find that the estimate for the parameter of risk aversion has a small upward bias, that the estimate for the subjective discount rate has a small downward bias, and that the estimate for the elasticity of intertemporal substitution is imprecise.

There are various ways in which we could improve the performance of our estimator. First, a natural way to proceed would be to characterize both the convergence properties as well as the large sample properties of the estimator. Second, we can improve the bandwidth selection for the bandwidths used in both the estimation of the continuation utility as well as the elements of the dynamic criterion function. Third, we can consider to use other conditional variables. Our analyses could be extended by considering another conditioning variable as well by including more than one conditioning variable by means of product kernels. For instance, long-run risk models already specify the relevant state variables, so this would be a natural extension. Fourth, it is known that a Nadaraya-Watson regression can generate a (small) downward bias in the estimated sequence of variable that defines the contraction mapping – which in our case is the continuation utility. The use of a local polynomial estimator could resolve this issue. Another advantage of local polynomial estimator is that it suffers less from boundary effects.

There is also room for other extensions. For instance, we have restricted ourselves to a single functional form of the class of Chew-Dekel preferences. However, other functional forms of the time-aggregator or the risk aggregator can be estimated in a similar fashion by exploiting the flexibility of kernel regressions. To this end, future research could be directed at specifying other functional forms, be they existing in the literature or not, and confront them with the data as well. A particular advantage of our approach is that the estimator permits convexity in the risk aggregator, which are present in, for instance, the generalized disappointment aversion functional. A subsequent comparison of the performance of various functional forms within the Chew-Dekel class might turn out to be interesting.
References


A Proofs

Proof of Proposition 1. Let us assume that \( \lim_{T \to \infty} v_T = \bar{v} \). Then,

\[
\bar{v}^{1 - \frac{1}{\psi}} = 1 - \beta + \beta \mathbb{E}_t \left[ \exp \left\{ \Delta \ln(C_{i+1})(1 - \gamma) \right\} \right]^{1 - \frac{1}{\psi}} \bar{v}^{1 - \frac{1}{\psi}} = 1 - \beta + \beta \exp \left\{ (1 - \gamma)\mu + \frac{1}{2}(1 - \gamma)^2\sigma^2 \right\}^{1 - \frac{1}{\psi}} \bar{v}^{1 - \frac{1}{\psi}} = 1 - \beta + \beta \exp \left\{ (1 - 1/\psi)\mu + \frac{(1 - \gamma)(1 - 1/\psi)}{2}\sigma^2 \right\} \bar{v}^{1 - \frac{1}{\psi}} = 1 - \beta + \xi \bar{v}^{1 - \frac{1}{\psi}},
\]

where we went to (A.2) by noting that i) \( \mathbb{E}_t[x_{i+1}] = \mathbb{E}[x] \) and, subsequently, that ii) \( \mathbb{E}[\exp(x)] = \exp \left\{ \mathbb{E}[x] + \frac{1}{2} \text{var}(x) \right\} \) when \( x \sim \mathcal{N} \). Some rewriting gives the result:

\[
\frac{1}{\xi} = \frac{\bar{v}^{1 - \frac{1}{\psi}}}{\bar{v}^{1 - \frac{1}{\psi}} - (1 - \beta)},
\]

\[
1 - \xi = \frac{1 - \beta}{\bar{v}^{1 - \frac{1}{\psi}} - (1 - \beta)},
\]

\[
\bar{v}^{1 - \frac{1}{\psi}} = \frac{\xi(1 - \beta)}{1 - \xi} + (1 - \beta) = \frac{1 - \beta}{1 - \xi},
\]

where we went to (A.6) by subtracting 1 from both sides, and the result follows from raising both sides of (A.7) to the power \( 1/(1 - 1/\psi) \).

B Genetic Search algorithms

This section heavily relies on the section in Heer and Maussner (2009). For details, please see Heer and Maussner (2009).

As the name suggests, the terminology of Genetic Search algorithms originates in biology. A set of candidate solutions is called the population, and the members of the population are labeled chromosomes. In our setting, one candidate for \( \hat{\theta}_T \) is a chromosome, and a population is composed of \( N \) chromosomes. One iterates in each step over new sets of candidate solutions. Of each chromosome, the fitness is evaluated, which is its ability to solve the problem at hand. In most problems, the fitness of a candidate is assigned a value by \( \mathbb{R}_+ \)-valued function \( f \) that takes on higher values for better solutions. In our case, chromosomes that return a low value for the criterion function, \( \delta \), have a high fitness.

The Genetic Search algorithm proceeds as follows. The chromosomes of the initial population are drawn from a random distribution. We can do so using a pseudo-random number generator. If there are no restrictions on, for instance, the parameter values, we can still use a random number generator but assign a very low fitness to chromosomes that do not satisfy the restrictions or even discard such chromosomes. When we have obtained the fitness for all chromosomes in a population, we generate offspring. Offspring is a new population of chromosomes, that are generated by the previous population.
The new chromosomes are combinations of (typically two) chromosomes of the previous population. These two chromosomes that generate a new chromosome are called parents. There are various methods to choose parents (see e.g. Heer and Maussner, 2009, p. 636–638), but these various methods are all in the spirit of natural selection. Suppose that a chromosome $i = 1, 2, ..., n$ has fitness $f(i)$ and relative fitness $p(i) \propto f(i)/\sum_{j=1}^{n} f(j)$. Then, the probability that a chromosome $i$ is selected as a parent is proportional to $p(i)$. A new chromosome is a combination of the parents’ chromosomes. There are various algorithms that determine what elements of which of the parent’s chromosome are carried over. These algorithms are labeled crossover algorithms.

However, as in nature, there are mutations; sudden changes in the genetic code due to copying mistakes or external forces. There are various ways to generate mutations, and we follow the most popular approach suggested by Michalewicz (2013). Let $C_j$ denote the $j$th elements in a child’s chromosome and $\tilde{C}_j$ denote a mutated element. Then, given a binary operator $l$, a $C_j$ mutates with probability $\pi$:

$$C_j' := \begin{cases} 
C_i + \lambda(r, t; b, T), & l = 0, \\
C_i - \lambda(y, r, t; b, T), & l = 1,
\end{cases}$$

where $\lambda(y, r, t; b, T) := y(1 - r(1 - (t/T)^{\beta})), \quad \beta := 1$.

We implement our strategy as follows. We draw chromosomes for the initial population from a normal distribution with mean zero and variance $\sigma^2$. We choose four parents at random (with replacement) from the preceding generation. Then, with probability 0.95 the best two will have two children. Mutations are given by (B.9) and the probability of mutations is given by

$$\pi(t; \mu) := \mu_1 + \frac{\mu_2}{t},$$

where $\mu_1$ and $\mu_2$ are positive scalars. We furthermore use $\mu_1 = 0.50$, $\mu_2 = 0.33$, and $\mu_3 = 2$. Since our chromosomes have to be from $\Theta$, we take the exponents of the chromosomes (to ensure they are positive), and we assign a negative value to the fitness when they are not within $\Theta$ so that these candidates are discarded by the algorithm. We use $\sigma = 1$. This method ensures relatively fast convergence with little risk of ending up at a local optimum. Since the problem is rather complex, we have to generate large populations and many generations (Heer and Maussner, 2009). Specifically, we generate $N = 1000$ populations and $T = 500$ generations. We rely on three crossover algorithms – arithmetic crossover, single-point crossover, and shuffle crossover – and we randomly select, with equal probability, one of the methods for each new generation.\(^{13}\)

\(^{13}\)See (Heer and Maussner, 2009, p. 638–639) for the definitions of these crossover algorithms.