A Powerful Test Needs to Be Size-Correct: Response to *"Robust Inference for Consumption-based Asset Pricing with Power."*

Frank Kleibergen[*] Zhaoguo Zhan[†]

November 15, 2022

Abstract

Statistical tests need to be size-correct, i.e., their rejection frequencies under the null hypothesis should not exceed the nominal significance level, before we can talk about their power. It is well established in the weak identification literature that commonly used t-tests (such as the Fama-MacBeth/Shanken and GMM t-tests) exhibit size distortion when identification conditions are at risk, while identification-robust tests remain size-correct. Furthermore, this literature has also produced tests that are both size-correct and optimal in terms of power. Therefore, these robust tests should be recommended over t-tests, and not vice versa.

Kleibergen and Zhan (2020) propose two asset pricing tests in the beta representation of expected asset returns for settings where the number of time-series observations does not greatly exceed the number of assets in the cross-section, because of which the usual asymptotic distributions of the traditional tests do not apply. The first test examines the full rank condition of the beta matrix, which is required for identifying risk premia; the second test (labelled as GRS-FAR) on the risk premia can be inverted to construct confidence sets. The tests are straightforward extensions of the Gibbons, Ross, and Shanken (GRS, 1989) test, and they are robust in the sense that their sizes, or rejection frequencies under the null hypothesis, do not depend on the statistical quality of the risk factors and the number of

[*] Email: f.r.kleibergen@uva.nl. Amsterdam School of Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands.

[†] Email: zzhan@kennesaw.edu. Department of Economics, Finance, and Quantitative Analysis, Coles College of Business, Kennesaw State University, GA 30144, USA.

time-series observations compared to the number of assets in the cross-section, both of which are empirically relevant issues in consumption-based asset pricing.

The robust tests of Kleibergen and Zhan (2020) are part of the now well-established literature on weak identification; see, e.g., Staiger and Stock (1997), Dufour (1997), Stock and Wright (2000), Kleibergen (2002, 2005, 2009), Moreira (2003), Andrews et al. (2006), Andrews and Cheng (2012), Beaulieu et al. (2013), Andrews (2016), Andrews and Mikusheva (2016a, b), and Andrews et al. (2019). The earlier papers in this literature show that when the identification strength of the parameters of interest (such as the risk premia) is weak, traditional tests, such as the commonly used t-test, exhibit size distortion, so they become unreliable. The later papers provide a variety of so-called weak identification robust inference methods which, unlike the t-test, remain size-correct regardless of the strength of identification. Moreover, they show that overcoming the size distortion of the traditional tests does not result in a loss of power, since some of the proposed procedures have provenly optimal power properties for both strongly and weakly identified settings; see, e.g., Andrews et al. (2006). Weak identification robust inference methods thus come at no cost in terms of power and are reliable for all identification strengths, so they are recommended for empirical applications.

Kroencke (2022) criticizes that the robust tests of Kleibergen and Zhan (2020) appear to lack power in small samples. At the heart of his critique sits, however, the opinion that the Fama-MacBeth (FM) estimator and FM/Shanken t-statistic on the risk premia remain well behaved when the betas are small. This opinion, which is motivated by just a few simulation exercises, goes against the well-established literature on weak identification which has pervasively shown that the distributions of such estimators and t-statistics become non-standard for settings where the betas are small.

We provided a detailed referee report on an earlier version of Kroencke (2022) with constructive comments, which, however, have been largely ignored. It is therefore not productive to do so again, and we just provide some main takeaways:

1 Size vs. power

It is worth noting that the two tests of Kleibergen and Zhan (2020) are related to each other. In particular, if the rank test indicates that the full rank condition of the beta matrix is not satisfied, then risk premia are considered unidentified in the beta representation. Consequently, confidence sets of the risk premia that result from inverting the GRS-FAR test will be unbounded, which reflects that no value of the risk premia can be rejected when there is no identification. This cannot happen when using t-tests, such as the FM/Shanken t-test. Even when we cannot reject a reduced rank value of the beta matrix, so we cannot reject that risk premia are not identified, the t-test will produce bounded confidence sets indicating that the risk premia are identified. The t-test does therefore not meet the requirement proven in Dufour (1997) that, if the tested parameter can be unidentified, a size-correct test procedure must have a positive probability of producing an unbounded confidence set. Hence, since t-tests, such as the FM/Shanken t-test, cannot generate unbounded confidence sets, they cannot be size-correct when used for testing a parameter which can be unidentified.

It serves no purpose to talk about power of tests if such tests do not control the size for all settings. For the FM/Shanken t-test on the risk premia, it has been well documented that its size can be distorted. This size distortion should not be surprising, since the t-test depends on the distribution of the risk premia estimator becoming normal when the sample size gets large, which is at risk when the strength of identification is poor; see, e.g., Kleibergen (2009). Moreover, the FM/Shanken t-test does not account for model misspecification, so it tends to over-reject the hypothesized risk premia; see, e.g., Kan, Robotti, and Shanken (2013).

Kroencke (2022) shows that the FM/Shanken t-test does not appear to be too bad in a few simulation experiments. The question is if these simulations are representative for all cases. To answer this question, a detailed analytical analysis has to be conducted, since only such an analytical approach is able to provide the full generality to cover all settings. These analytical analyses have been conducted in the weak identification literature, see, e.g., Staiger and Stock (1997), Stock and Wright (2000), and Kleibergen (2009), and all of these have shown that the distributions of estimators, like, the FM two-pass risk premia estimator, become non-standard when the identification conditions for the parameters start to fail. Tests using these estimators therefore become unreliable when the identification of the parameters becomes questionable.

Put differently, while the FM/Shanken t-test may appear powerful so it tends to support risk factors in empirical studies, this t-test is not trustworthy so it has to be considered with caution, since it does not produce unbounded confidence sets even if risk premia are unidentified; see Dufour (1997). The resulting over-rejection of the t-test has motivated weak identification robust tests on risk premia, including those proposed by, e.g., Beaulieu, Dufour, and Khalaf (2013), Kleibergen and Zhan (2020). Unlike the commonly used t-test, these weak identification robust tests remain size-correct regardless of the strength of identification.

2 Rank of the beta matrix vs. univariate beta

The rank test proposed by Kleibergen and Zhan (2020) examines the beta matrix for the general setup of N test assets and K risk factors for which the dimension of the beta matrix is N by K. To identify the risk premia, a full rank $N \times K$ beta matrix is needed. The rank test of Kleibergen and Zhan (2020) thus needs to be passed in order to achieve identification of the risk premia in cross-sectional regressions.

In contrast, Kroencke (2022) proposes a univariate test on a single beta for evaluating the pricing ability of risk factors, so the cross-sectional dimension is reduced to $N = 1$ with $K = 1$. Does passing such a univariate test imply identification of the risk premia in cross-sectional regressions? It appears not. From an econometric perspective, the univariate test and the rank test are not comparable with respect to size and/or power: since the hypothesis of interest differs, they can not replace one another.

3 Misspecification

Kleibergen and Zhan (2020) are explicit that the GRS-FAR test tests both misspecification and the parameter of interest (i.e. risk premia), so it is not informative about the parameter of interest in case of misspecification. Page 535 of Kleibergen and Zhan (2020) states: “In the case of misspecification, the factor pricing restrictions no longer fully apply, so we need to adapt our inference methods, adopting misspecification-robust methods; see, for example, Kan, Robotti, and Shanken (2013) and Gospodinov, Kan, and Robotti (2014, 2018)."

Nevertheless, Kroencke (2022) criticizes the GRS-FAR test for not being useful for inferring risk premia when a model is misspecified. Kroencke (2022)'s suggested alternative, the FM/Shanken *t*-test does, however, not explicitly account for misspecification; see, e.g., Kan, Robotti, and Shanken (2013). To jointly account for both misspecification and weak identification of risk premia, a double robust test is needed; see, e.g., Kleibergen and Zhan (2021).

4 Validity of the bootstrap

Kroencke (2022) proposes the bootstrap for constructing confidence sets of the risk premia and the relative rate of risk aversion. The bootstrap does, however, not always provide a valid manner to compute a confidence set. The bootstrap is generally valid to compute confidence sets when using, so-called, asymptotically pivotal statistics, whose limiting distributions do not depend on other unknown (nuisance) parameters; see Horowitz (2001). The limiting distribution of the risk premia estimator depends on the rank value of the beta matrix. A different (non-normal) limiting distribution results when the beta matrix is of reduced rank, compared to when the beta matrix is of full rank. This shows why the risk premia estimator is not asymptotically pivotal. The bootstrap uses the sample analog of the beta matrix which is by construction always of full rank, so the bootstrap cannot approximate the distribution of the risk premia estimator when the true beta matrix is of reduced rank. This reasoning similarly applies to the estimator of the relative rate of risk aversion, and explains why the bootstrap is not reliable in our context of interest.

Another argument why the bootstrap is not valid for constructing confidence sets of the risk premia (and similarly, the relative rate of risk aversion) is that it always leads to bounded confidence sets even when we cannot reject the hypothesis of not identified risk premia, which corresponds with a lower rank value of the beta matrix.

5 Nonlinear GMM-AR and rank tests

For nonlinear GMM, Kleibergen and Zhan (2020) just employ the GMM-AR test advocated by Stock and Wright (2000) and a GMM rank test from Wright (2003). The GMM-AR test is a so-called identification robust test so, like other identification robust tests, its limiting distribution is not affected by the identification or Jacobian rank assumption on the analyzed parameters. All these identification robust tests can produce unbounded confidence sets which then reflect identification issues.

Kroencke (2022) criticizes that the GMM-AR and GMM rank tests have numerical issues in small samples, and proposes the GMM t-test for inferring the relative rate of risk aversion. The GMM t-test, like the FM/Shanken t-test, is not identification-robust, so it does not produce unbounded confidence sets or lead to valid inference under weak identification; see, e.g., Stock and Wright (2000).

6 Conclusions

At a glance, robust tests may not always appear as powerful as the FM/Shanken and GMM t-tests. They, however, remain size-correct while t-tests are often not. The power of t-tests resulting from specific simulation experiments is therefore not very revealing. For this reason, we propose robust tests, since its power is easily interpretable as the size is controlled. When resorting to limiting distributions, which Kleibergen and Zhan (2020) refrain from since the number of time-series observations is too close to the number of assets in the cross-section to apply them, specific robust tests are also optimal so they dominate the t-tests both in terms of size and power. The usage of traditional tests, like the FM t-test, is an important reason for the existence of a zoo of priced factors in asset pricing. The risk premia of many of these factors are, however, quite weakly identified, so it is natural to gauge them using identification robust tests. These tests show that the risk premia of many of the risk factors might very well be non-identified, so usage of identification robust tests disciplines the factor zoo.

References

Andrews, D.W.K. and X. Cheng, Estimation and inference with weak, semi-strong and strong identification, *Econometrica,* 2012, **80**, 2153-2211.

Andrews, D.W.K., M.J. Moreira and J.H. Stock, Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression, *Econometrica,* 2006, **74**, 715-752.

Andrews, I., Conditional Linear Combination Tests for Weakly Identified Models, *Econometrica,* 2016, **84**, 2155-2182.

Andrews, I. and A. Mikusheva, A Geometric Approach to Nonlinear Econometric Models, *Econometrica,* 2016a, **84**, 1249-1264.

Andrews, I. and A. Mikusheva, Conditional inference with a functional nuisance parameter, *Econometrica,* 2016b, **84**, 1571-1612.

Andrews, I., J. Stock and L. Sun, Weak instruments in IV Regression: Theory and Practice, *Annual Review of Economics*, 2019, **11**, 727-753.

Beaulieu M. C., Dufour J. M., Khalaf L., Identification-Robust Estimation and Testing of the Zero-Beta CAPM, *Review of Economic Studies*, 2013, **80**, 892-924.

Dufour, Jean-Marie, Some impossibility theorems in econometrics with applications to structural and dynamic models, *Econometrica,* 1997, **65**, 1365-1387.

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken, A test of the efficiency of a given portfolio, *Econometrica,* 1989, **57**, 1121-1152.

Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti, Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors, *Review of Financial Studies,* 2014, **27**, 2139-2170.

Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti, Asymptotic variance approximations for invariant estimators in uncertain asset-pricing models, *Econometric Reviews,* 2018, **37**, 695-718.

Horowitz, J.L, The bootstrap, *Handbook of Econometrics,* Volume 5, Elsevier Publisher, 2001.

Kan, R., C. Robotti and J. Shanken, Pricing Model Performance and the Two-Pass Cross-Sectional Regression Methodology, *Journal of Finance,* 2013, **68**, 2617-2649.

Kleibergen, F., Pivotal Statistics for testing Structural Parameters in Instrumental Variables Regression, *Econometrica,* 2002, **70**, 1781-1803.

Kleibergen, F., Testing Parameters in GMM without assuming that they are identified, *Econometrica,* 2005, **73**, 1103-1123.

Kleibergen, F., Tests of risk premia in linear factor models, *Journal of Econometrics,* 2009, **149**, 149-173.

Kleibergen, F. and Z. Zhan, Robust Inference for Consumption-Based Asset Pricing, *Journal of Finance,* 2020, **75**, 507-550.

Kleibergen F. and Z. Zhan, Double Robust Inference for Continuous Updating GMM, Working paper, 2021, arXiv:2105.08345.

Kroencke, Tim A., Robust Inference for Consumption-based Asset Pricing with Power, *Critical Finance Review,* 2022.

Moreira, M.J., A Conditional Likelihood Ratio Test for Structural Models, *Econometrica,* 2003, **71**, 1027-1048.

Staiger, D. and J.H. Stock, Instrumental variables with weak instruments, *Econometrica,* 1997, **65**, 557-586.

Stock, J.H. and J. Wright, GMM with weak identification, *Econometrica,* 2000, **68**, 1055-1096.

Wright, Jonathan H., Detecting lack of identification in GMM, *Econometric Theory,* 2003, **19**, 322-330.