With big data come big problems: pitfalls in measuring basis risk for crop index insurance

Matthieu Stigler[1], Apratim Dey[2], Andrew Hobbs[3], and David Lobell[1]

[1]Center on Food Security and the Environment, Stanford University, USA
[2]Department of Statistics, Stanford University, USA
[3]Department of Economics, University of San Francisco, USA

November 30, 2021

New satellite sensors will soon make it possible to estimate field-level crop yields, showing a great potential for agricultural index insurance. This paper identifies an important threat to better insurance from these new technologies: data with many fields and few years can yield downward biased estimates of basis risk, a fundamental metric in index insurance. To demonstrate this bias, we use state-of-the-art satellite-based data on agricultural yields in the US and in Kenya to estimate and simulate basis risk. We find a substantive downward bias leading to a systematic overestimation of insurance quality.

Our results indicate that more research will be needed to fully leverage new technologies in index insurance. As a first step in this direction, we derive a new theorem characterizing the estimation bias. Our formula accurately approximates the empirical bias simulated from the satellite data, and provides a useful tool for practitioners to quantify bias in insurance quality.

Keywords: index insurance, agricultural risk, estimation bias, high-dimensional data, satellite technology

1 Introduction

Index insurance is a promising tool to reduce the risk faced by smallholder farmers. By linking payouts to a regional index instead of individual losses, it reduces moral hazard, adverse selection and transaction costs. But delinking payouts from individual losses creates *basis risk*, the possibility that a farmer experiences a loss yet does not receive any indemnity. Basis risk substantially reduces the benefit of index insurance, and if severe enough can make it worse than no insurance at all. Further, Clarke (2016), Carter et al. (2017), and others argue that basis risk is likely among the most important barriers to index insurance adoption, and that the basis risk of some index insurance schemes has been be very high (Clarke et al., 2012; Jensen et al., 2016).

Recent improvements in satellite remote sensing and machine learning show great potential to improve the accuracy of index and thus reduce basis risk. These technologies have triggered a very active literature extending far beyond the field of economics.[1] These technologies have led to three major shifts in the design of agricultural indices. First, satellite data has helped to design better weather-based indices: while early products were based on local weather stations, satellite data has increased the spatial resolution of weather indices, and facilitated the incorporation of new variables such as soil moisture. Second, it has led to a shift

[1]See for example the special issue of *Remote Sensing* on "Earth Observation for Index Insurance" 2021, 13(5), or reviews by Benami et al. (2021) and De Leeuw et al. (2014).

from input-based weather indices towards higher-accuracy output-based indices based on vegetation indices observed with optical satellite sensors. In these two first approaches, the satellite data is used primarily to obtain a better index, while the assessment of the quality of the index is conducted using traditionally-collected field-level yield data. In the third approach, satellite data is used directly to estimate farm or field-level crop yields. Although accurately predicting individual yields currently remains a challenge, rapid progress is being made and this approach shows great potential for deriving very accurate output-based indices. Even more importantly, it will provide a cost-effective way to assess the quality of a given index over a large number of fields, helping insurers to design better insurance zones and making it easier for governments, researchers, and others to reliably assess the quality of index insurance products.

This paper identifies and studies a new challenge associated with assessing index insurance quality with more granular data: existing measures of quality are biased in large-N (number of farms), small-T (number of time periods) samples. We show that R^2-derived estimates of basis risk are systematically biased downward in the small-T large-N case, meaning practitioners who do not take this into account will generally overestimate the quality of index insurance products. Intuitively, this bias arises from the fact that the basis risk estimates are functions of the covariance matrix between fields. The covariance matrix between fields has $N \cdot (N-1)/2$ parameters, yet only $N \cdot T$ observations to estimate it. Having more fields N than time periods T is very typical of agricultural data, and is going to be exacerbated by developments of satellite data methods, which are particularly suited to extend the sample over space, but are unfortunately only available for a few recent years. The resulting bias has gone unnoticed in the existing literature and has important implications for both past and future estimates of basis risk.

After documenting the bias in various measures of basis risk, we analyze it theoretically. We focus on linear measures of basis risk, which allows us to connect our problem with a rich literature in statistics. We start with a review of models to parameterize the inter-field covariance matrix, focusing on the *spiked* model introduced by Johnstone (2001). We discuss then how the high-dimension, low-sample-size (HDLSS) framework introduced by Hall et al. (2005) can help us understand the bias of linear basis risk measures. In the HDLSS framework, the sample size T is assumed fixed, while the number N of variables is assumed to grow to infinity. This corresponds exactly to the situation we are facing in index insurance, where satellite data techniques increase sample sizes over space much faster than over time. Using results from the HDLSS literature, we provide a new theorem deriving the theoretical bias of our linear measure of basis risk. Going back to the simulations, we find that our theory predicts the empirical bias remarkably well.

The bias we study in this paper is particularly pernicious for two reasons. First, we show the bias in basis risk measurement can actually be worsened by higher resolution data. Second, the bias is greatest when individual yields are poorly correlated, meaning it is likely to be particularly severe in smallholder systems in developing countries, which is precisely where new satellite data promise to have the largest positive impact. In light of these findings, understanding this bias is essential to realizing the promise of high resolution satellite data for index insurance.

This paper's findings also provide a potential explanation for observed low uptake of existing index insurance products.[2] Farmers are experts on their yields and are aware of how they relate to their neighbors, and likely have an accurate understanding of how correlated their yields are to their neighbors'. Since the bias we study in this paper is essentially a result of inaccurate estimates of inter-field correlations due to a small number of observed time periods, farmers who have more accurate understandings of these correlations ought to buy insurance less often than biased estimates would suggest.

2 Measures of basis risk

Index insurance products are usually assessed following two broad approaches. In the first one, the interest is on basis risk, which is a measure of the frequency and size of errors in predicting individual yields and/or harvest losses. The second seeks to evaluate how the insurance product derived from this index performs. This entails typically specifying a indemnity and premium functions and assuming a utility function for the

[2]See Carter et al. (2014) and Carter et al. (2017) for discussions of the literature on low index insurance uptake

farmer. In this paper, we focus on the first approach. Basis risk is sometimes defined as the probability of a farmer experiencing a loss yet not receiving an indemnity. As such, it can be estimated as a simple conditional probability, the false negative probability. However, this measure fails to capture the severity of the prediction error, which has important consequences for farmer welfare. For this reason, Elabed et al. (2013) suggest focusing on the field-level share of the variation in yields not explained by the index. This is equivalent to using $1 - R_i^2$, where R_i^2 is the coefficient of determination between the yields of field i and the index.

In this paper, we focus on output-based indices such as the zone average yield. This comes from our initial motivation to assess the benefits of third-generation datasets, which (will) allow estimates of yields at the field-level. That said, the same results hold for traditional area yield insurance contracts; the source of the field-level yield estimates does not matter. In recent work on output-based indices, Stigler and Lobell (2021) discuss how to aggregate the field-specific R_i^2 and propose a zone-specific *total* $\overline{\overline{R^2}} \equiv 1 - \sum_i SSR_i / \sum SST_i$. This $\overline{\overline{R^2}}$ is a generalization of the individual R^2 written as $R_i^2 \equiv 1 - SSR_i/SST_i$, where SSR and SST stand respectively for sum of squared residuals and total sum of squares. This measure is simply a variance-weighted average of the individual R_i^2, meaning it puts more weight on farmers who are exposed to more risk. The $\overline{\overline{R^2}}$ measure can alternatively be obtained by running N field-specific regressions of the index on individual yields and aggregating their SSR_i and SST_i into $\overline{\overline{R^2}} \equiv 1 - \sum_i SSR_i / \sum SST_i$. In the case of an output-based area-yield index, Stigler and Lobell (2021) propose an alternate, numerically identical, formula:

$$\overline{\overline{R^2}} = tr\left(\Sigma \mathbf{1}(\mathbf{1}'\Sigma\mathbf{1})^{-1}\mathbf{1}'\Sigma\right)/tr(\Sigma) \tag{1}$$

Here, Σ is the covariance matrix between individual fields, and $\mathbf{1}$ denotes a vector of 1, of dimension N. While numerically equivalent to field-specific regression, formula (1) has the advantage of establishing the connection between the basis risk and the covariance of fields. Intuitively, the strength of the index depends on the strength of the off-diagonal elements: a diagonal covariance matrix (uncorrelated fields) would result in higher basis risk than a covariance matrix with many positive off-diagonal elements (many correlated fields).

Stigler and Lobell (2021) show also that the formula (1) can be generalized to a broader class of output-based indices, which use field-specific weights to form the index, $f_t = \sum_i w_i y_{it}$, or in matrix form, $f = Y\mathbf{w}$. The area-yield index is a special case in this class, with $\mathbf{w} = \mathbf{1}/N$. The formula becomes then:

$$\overline{\overline{R^2(\mathbf{w})}} = tr\left(\Sigma \mathbf{w}(\mathbf{w}'\Sigma\mathbf{w})^{-1}\mathbf{w}'\Sigma\right)/tr(\Sigma) \tag{2}$$

They show that this quantity is not maximized using the area-yield index $\mathbf{1}/N$, but instead taking the first principal component (PC) of the covariance matrix Σ, $\mathbf{w}^\star = PC_1(\Sigma)$. Evaluated at this optimal $\mathbf{w}^\star$, the objective function $\overline{\overline{R^2(\mathbf{w}^\star)}}$ turns out to be equal to the share of the first eigenvalue of Σ, that is $\overline{\overline{R^2(\mathbf{w}^\star)}} = \lambda_1 / \sum \lambda$. The $\overline{\overline{R^2(\mathbf{w}^\star)}}$ is an interesting measure that defines the upper-bound any index can achieve (according to the total R^2 criterion) for a given zone. In that sense, it can be interpreted as a measure of zone quality, a low $\overline{\overline{R^2(\mathbf{w}^\star)}}$ for a given zone indicating that even the best index would not perform very well. The connection between $\overline{\overline{R^2(\mathbf{w}^\star)}}$ and the eigenvalues of Σ indicates that the $\overline{\overline{R^2(\mathbf{w}^\star)}}$ is equivalent to the usual definition of $\lambda_1 / \sum \lambda$ in terms of the *percentage of total variance captured by the first principal component*. In addition, and of particular relevance for the current paper, the statistical properties of sample eigenvalues are a very well-studied problem in statistics.

Admittedly, linear correlation measures have limitations in the context of index insurance. Arguably, it is more important for an index to accurately predict yield losses than to predict good harvests. Various approaches have been suggested to take this into account, ranging from quantile regression (Conradt et al., 2015) to more sophisticated left-tail dependence indices (Bokusheva, 2018). In the following, we include also a quantile version of our total R^2 measure, based on the quantile pseudo R^2 developed by Koenker and Machado (1999). Koenker and Machado (1999) suggest a pseudo $R^2(\tau) = 1 - V(f,\tau)/V(const,\tau)$ at quantile τ, where $V(\tau)$ is the quantile analogous to the SSR. In a similar way to our total R^2, we define

the total quantile pseudo R^2 as $R^2_q = 1 - \sum V_i(f,\tau)/\sum V_i(const,\tau)$, and use the value of $\tau = 0.3$ following previous literature (Bucheli et al., 2020). The bias we identify in this study may also apply to other nonlinear measures of basis risk, but we leave that for future studies.

3 Data and empirical simulations

3.1 Data

We use two state-of-the-art datasets of satellite-estimated yields in the USA and in Kenya to illustrate the potential bias in basis risk measures. Both datasets contain maize yield predictions produced with the Scalable Yield Mapper (SCYM) model initially developed by Lobell et al. (2015). The SCYM model is one of the most advanced yield prediction models available to date (see Jin et al., 2017b; Azzari and Lobell, 2017; Deines et al., 2021 for the US and Burke and Lobell, 2016; Jin et al., 2017a, 2019; Lobell et al., 2020 for Sub-Saharan Africa), and has been already used to analyze various questions such as the effect of cover crops, of conservation tillage or the dynamics of crop expansion (Seifert et al., 2018; Deines et al., 2019; Stigler, 2018). The dataset has been used specifically for analyzing crop insurance in the US in Stigler and Lobell (2020) and in Kenya in Stigler and Lobell (2021).

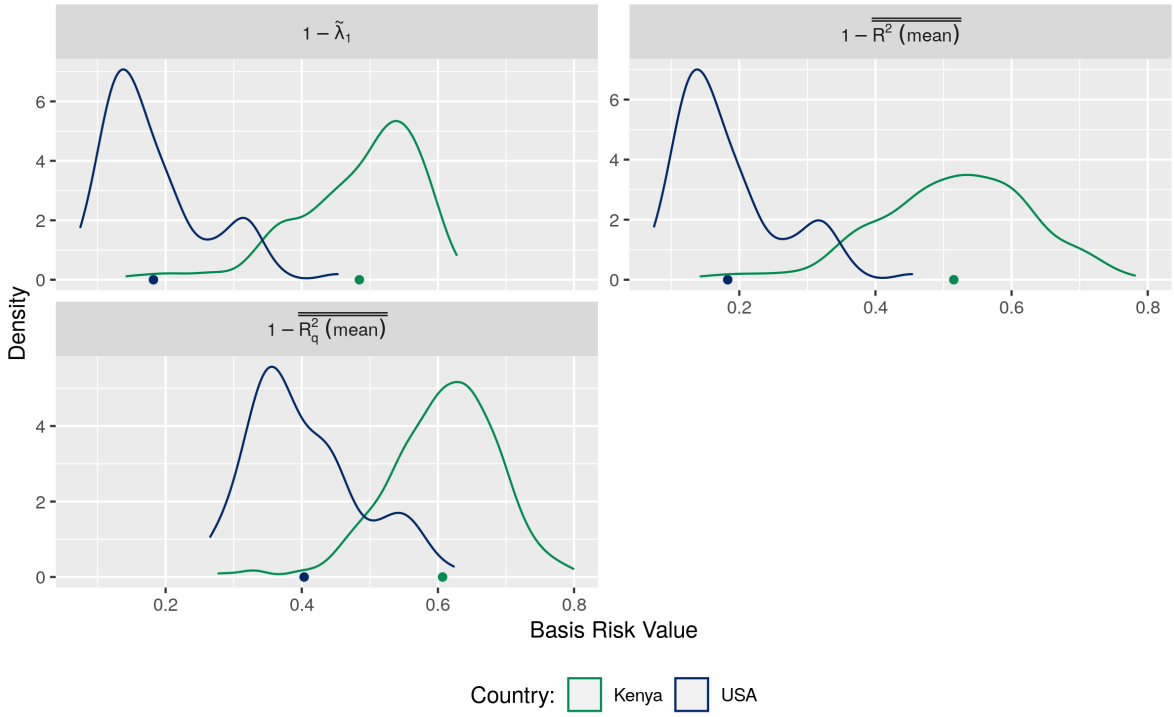
While the US and Kenya datasets share a common methodology, they also differ in several respects. First, we have twenty years of data for the US: from 2000 to 2019, while for Kenya we have only four: from 2016 to 2019. This difference is due to the fact that fields are much smaller in Kenya which means that higher resolution satellite images are required (from Sentinel-2), and those images are only available for recent years. Another difference lies in the fact that maize is mainly cultivated in rotation together with soybeans in the US, while this is less common in Kenya. For the US, this means we observe a large number of missing maize values for those fields practicing rotation, which makes estimating accurate covariance matrices more difficult. We adopt a simple solution, and focus on the fields that only cultivate corn over the 2000-2019 period, and select counties that have at least 30 observations. Doing so, we end up with a sample of 37 *zones* for the US. For Kenya, almost all fields cultivate maize every year, and we randomly sample 200 fields in each of 453 Kenyan sub-counties (Kenya's smallest administrative unit). A last aspect where both datasets differ is in the quality of the satellite predictions. Predictions are typically much better in the US, characterized by large uniform fields, than in Kenya, which has smaller fields and more heterogeneous cultivation practices. Deines et al. (2021) report that the SCYM yield estimates in the US have a r^2 of 0.45 when compared to a ground-truth dataset at the field level, raising to 0.69 when assessed against county-level means instead. On the other hand, Jin et al. (2019) report that the yield estimates in Kenya have an agreement of about 50% against district means. Clearly, the current accuracy of the satellite-based yield predictions is not yet perfect, and more research is still needed before using these datasets at an operational level for routine insurance assessment. Witnessing the rapid progress made in the field in the last ten years, there is good reason to believe that in a near future yield estimates will be much more accurate. In this paper, we focus on another source of error for insurance applications, that due to the small-T large-N setting, that has largely gone unnoticed in the literature.

3.2 Basis risk measures in SCYM data

Our analysis here proceeds in two steps. We first estimate various basis risk measures using both SCYM data sets. We then estimate the bias associated to these measures. To do so, we employ a Monte Carlo (MC) simulation experiment using pseudo true values calibrated to the data at hand. More precisely, we pretend that the covariance matrices and the basis risk measures estimated in the first step are the true ones, and simulate random samples of various T sizes using these pseudo-true covariances. We then re-estimate the basis risk measures on the simulated samples, and infer the bias by comparing those simulated values to the pseudo-true basis risk measures used to simulate the data.

Starting with the initial estimates of basis risk, we focus on three measures, 1) the total R^2 using the county mean as the index, 2) the total quantile pseudo R^2 using also the county mean, and 3) the total R^2

Figure 1: Basis risk measure on SCYM data



using the optimal index, which is equivalent to the share of the first eigenvalue. Figure 1 shows the results for these three measures, highlighting a stark difference between Kenya and the USA. Remembering that basis risk is defined as $1 - R^2$, we see that the basis risk is much lower in the USA than in Kenya according to the three measures.[3] This result is consistent with the structure of agricultural production in the two settings, large-scale farms in the US use relatively similar production technologies, whereas smallholder farmers Kenya are very heterogeneous. It is therefore unsurprising that basis risk, interpreted as the lack of correlation between fields, is much higher in Kenya than in the USA. However, these estimates are based on small-T samples (T=4 for Kenya, and T=20 for the USA), and we show in the next section that they underestimate basis risk as predicted by the theory.

3.3 SCYM-based simulations of the bias

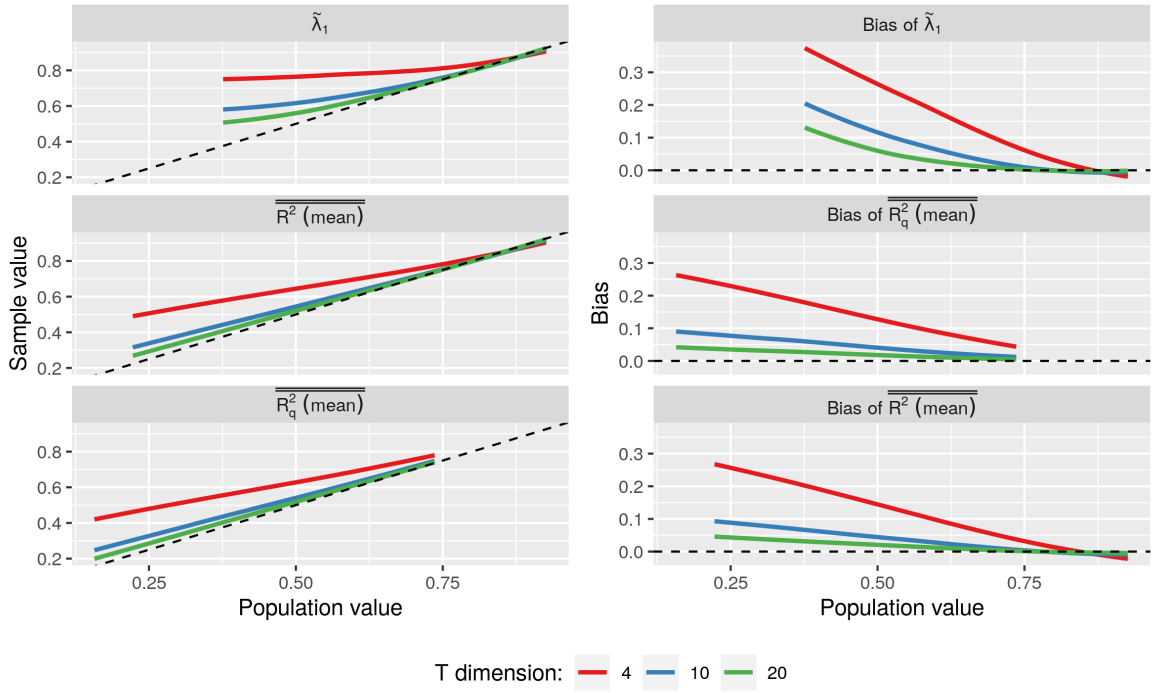
To gauge the reliability of the estimates obtained above, we proceed to a Monte Carlo simulation. Using the covariance matrices and means implicitly estimated above and pretending those are pseudo-true covariance matrices, we simulate random samples assuming a normal distribution. We do this for three different T dimensions representative of dataset found in practice, T=4, 10 and 20, and for each of these run 500 simulations. For each simulated sample, we recompute the three measures of basis risk, average them over the 500 samples, and compute the resulting bias by comparing these averages to the pseudo-true values obtained from the pseudo-true covariance matrices.[4]

Figure 2 shows the bias as estimated by simulation. Values on the x-axis denote the pseudo-true popula-

[3]Interestingly, these three measures of basis risk are highly correlated to each other, the lowest correlation being 92% between the quantile pseudo R^2 and the first eigenvalue.

[4]For the linear basis risk metrics, the population value is directly obtained from on the covariance matrix based on the formulas derived above. For other measures such as the quantile pseudo R^2, we don't have analytical formulas for the population values, and hence obtain them by simulating with a large sample of 25000 observations.

Figure 2: Bias from simulations based on SCYM



tion value, and on the y-axis the average sample values (left column) and bias (right column). We note a very similar phenomenon over each basis risk metric: the bias is relatively high for low values of the pseudo-true basis risk, and tend to decrease for increasing values of basis risk, with a possible sign reversal for very high values. This indicates that an insurer assessing the basis risk of a zone will tend to be over-optimistic about the R^2 measure, and hence under-estimate the basis risk itself. This upward bias in the R^2 is relatively large, and even larger in percentage terms, with an upward bias of 150% for low values of the quantile pseudo R^2 measure. The bias decreases for larger values of T, suggesting it would eventually disappear with a large enough sample over the T dimension. Taken together, the fact that the bias is higher for lower value and for lower T suggests that the bias in the initial estimates in Figure 1 is relatively modest for the USA (T=20 and high R^2 values observed), but much more important in Kenya (T=4 and low R^2 values observed).

The calibrated Monte Carlo exercise suggests that there is a substantive upwards bias in the total R^2 measure, resulting in downwards bias for the linear basis risk metrics considered. However, this approach of simulating pseudo-true covariance matrices faces several limitations. First, it rests on an initial estimate of the covariance matrix, which is itself potentially biased. Second, and most important, it is difficult to know whether these insights generalize to other cases, since the results were derived for specific covariance matrices. Generalizing these results calls for analyzing the bias theoretically, which is the subject of the next section.

4 Theory

The simulations in Section 3 show that there is potentially a substantive bias in various basis risk measures. These results were obtained by pretending that the sample covariance matrices we estimated are the population ones, and simulating from this population matrix. This is somewhat artificial, since the simulations show that our initial estimate could be biased. Escaping this conundrum calls for a more formal approach,

assuming a-priori a specific covariance matrix to be used as data-generating process (DGP). This makes it possible to simulate data by controlling and varying the *true* parameters, and assessing how sample estimates behave with respect to these known parameters.

4.1 Choice of the data generating process

To derive analytical results on the estimation of the basis risk metrics, we need to specify a data generating process. To do so, we adopt the so-called spiked covariance model introduced by Johnstone (2001). This model assumes that the eigenvalues of the covariance matrix are spiked, i.e. that a few eigenvalues clearly dominate the rest of the eigenvalues. This corresponds to the idea that a few common factors explain much of the variability of the data, and that the remaining variability is idiosyncratic. In fact, Wang and Fan (2017) show that a spiked structure can be generated from a factor model. This factor structure is exactly the assumption behind index insurance, which posits that a single common index will predict well individual yields. It is also implicitly assumed in existing literature on index insurance that models farm-level risk as a linear combination of covariate shocks and idiosyncratic farm-level shocks (See Miranda 1991, Mahul 1999 and Conradt et al. 2015 for example). This suggests that the spiked model is a very natural starting point as generating process for agricultural data, noting that the main determinants of yield (weather, input and output prices) are highly correlated within a zone. As another benefit, the spiked model is the subject of a vast theoretical literature (see for reviews Johnstone and Paul, 2018), providing us with important tools to understand the bias of our estimates.

The idea behind the spiked model is to specify the eigenvalues of the covariance matrix, assuming that a few eigenvalues are much larger than the other ones, and furthermore grow with the dimensionality N of the sample. In other words, letting $\lambda_1 > \lambda_2 > \ldots > \lambda_M > \lambda_{M_1} \geq \ldots \geq \lambda_N$ represent the ordered eigenvalues of Σ, one assumes that there are at most M *large* eigenvalues, and that the remaining $N - M$ are bounded. In the following, we will use the simplest specification, assuming that there is only one large spike $\lambda_1 = aN^\alpha$, and that the remaining eigenvalues are all equal to a constant b. Denoting Λ the diagonal matrix whose elements are the ordered eigenvalues, the population covariance matrix is then specified as $\Sigma = Q\Lambda Q^{'}$, where Q is an orthogonal matrix ($QQ^{'} = I$). Using an orthogonal matrix guarantees that the eigenvalues of Σ are the same as Λ, that is $[aN^\alpha, b, \ldots, b]$.

As discussed above, a measure of minimum zonal risk is given by $1 - \overline{\overline{R^2(w^\star)}} = 1 - \lambda_1 / \sum \lambda \equiv 1 - \tilde{\lambda}_1$, which indicates the minimum $1 - \overline{\overline{R^2(w)}}$ that can be reached within a zone by any linear index. From now on, we will focus on the population $\tilde{\lambda}_1 \equiv \lambda_1 / \sum \lambda$, and its sample counterpart, $\hat{\tilde{\lambda}}_1 \equiv \hat{\lambda}_1 / \sum \hat{\lambda}$. Starting with the population $\tilde{\lambda}_1$, in the spiked model it takes the value of $\tilde{\lambda}_1 = aN^\alpha / (aN^\alpha + (N-1)b)$. The behavior of $\tilde{\lambda}_1$ with growing N will depend on the value of α, and we distinguish three cases:

- Vanishing spike: $\alpha < 1$, and hence $\tilde{\lambda}_1 \xrightarrow{N \to \infty} 0$
- Constant spike: $\alpha = 1$, and hence $\tilde{\lambda}_1 \xrightarrow{N \to \infty} a/(a+b)$
- Expanding spike: $\alpha > 1$, and hence $\tilde{\lambda}_1 \xrightarrow{N \to \infty} 1$

Which spike structure should be considered is a complicated question. Typically, a researcher faces a dataset with given T and N, and asking the thought experiment of what would happen if she had an infinite number of fields N is somewhat abstract. One could eventually argue that when the dimension N is increased by extending the sample over space, adding fields further away would possibly reduce the value of $\tilde{\lambda}_1$, which would correspond to the vanishing spike case. Conversely, if the sample is extended by adding more pixels or fields within a zone, $\tilde{\lambda}_1$ could alternatively increase (expanding spike). At the theoretical level, the vanishing and expanding spike are, however, not very interesting since they only allow two extreme values of either 0 or 1. In the following, we focus hence on the constant spike, assuming that the value of $\tilde{\lambda}_1$ is constant, and equal to $a/(a+b)$. This will allow us to investigate the behavior of the basis risk measure for a variety of cases of $\tilde{\lambda}_1$, representing both low and high homogeneity zones that we encountered in Section 3.

4.2 Analytical results

Turning now to the behavior of the sample eigenvalue under the moderate spike, we need now to make assumptions on the ratio between T and N. There exist broadly three frameworks in statistics: 1) traditional asymptotics, with N fixed and $T \to \infty$ so that $N/T \to 0$, 2) random matrix theory $N/T \to c$ for a constant c and 3) high-dimension low-sample size (HDLSS), where $N/T \to \infty$. The latter case, HDLSS, appears the most appropriate to describe third-generation datasets, where T is considered fixed, and N is allowed to grow very large. The HDLSS was introduced by Hall et al. (2005), with notable contributions from Ahn et al. (2007); Jung and Marron (2009), see also Aoshima et al. (2018) for a review. While important results have been derived for the raw sample eigenvalue $\hat{\lambda}_1$ (see Ahn et al., 2007), there is, to the best of our knowledge, no result available on the relative eigenvalue $\hat{\tilde{\lambda}}_1 \equiv \hat{\lambda}_1 / \sum \hat{\lambda}$. To fill this gap, we derive a new result, describing the distribution and the bias of $\hat{\tilde{\lambda}}_1$, building on the seminal work by Ahn et al. (2007). In what follows, $\xrightarrow{d}$ denotes convergence in distribution and $\xrightarrow{p}$ denotes convergence in probability. See the appendix for details on the notation.

Theorem 1 (Distribution of the share of the first sample eigenvalue.)**.** *For $1 \leq t \leq T$, set $Y_t = (y_{1t}, ..., y_{Nt})$ and assume that $Y_t \overset{iid}{\sim} \mathcal{N}_N(\mu, \Sigma)$ where $\mathcal{N}_N(\mu, \Sigma)$ denotes the N-dimensional Normal distribution with mean vector $\mu \in \mathbb{R}^N$ and covariance matrix $\Sigma \in \mathbb{R}^{N \times N}$, and mu is unknown. Also assume a spiked model for Σ i.e. the eigenvalues of Σ satisfy $\lambda_1 = aN^\alpha$, and $\lambda_2 = \cdots = \lambda_N = b$ for some positive constants a, b. We have the following results.*

1. *Vanishing spike: $\hat{\tilde{\lambda}}_1 \xrightarrow{p} \frac{1}{T-1}$, noting that $\tilde{\lambda}_1 \to 0$.*
2. *Constant spike: $\hat{\tilde{\lambda}}_1 \xrightarrow{d} \frac{aC^2+b}{aC^2+b(T-1)}$ where $C^2 \sim \chi^2_{T-1}$, noting that $\tilde{\lambda}_1 \to \frac{a}{a+b}$.*
3. *Expanding spike: $\hat{\tilde{\lambda}}_1 \xrightarrow{p} 1$ noting that $\tilde{\lambda}_1 \to 1$.*

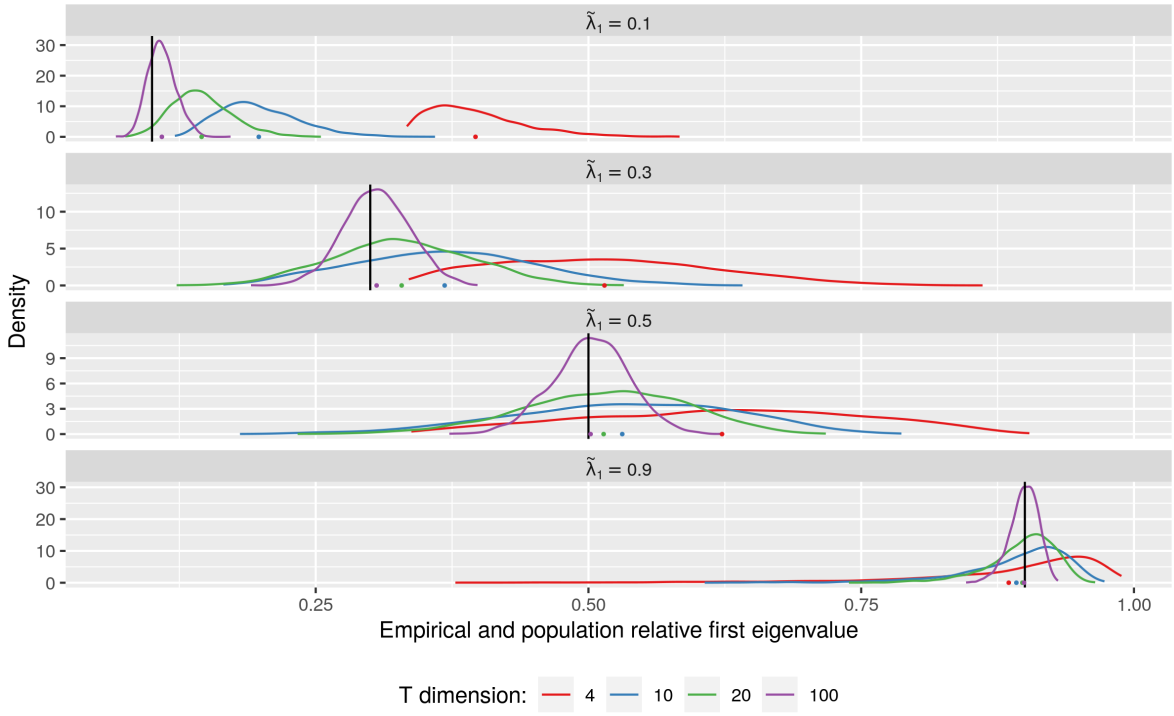
Theorem 2 (Bias of the share of the first sample eigenvalue.)**.** *Assume the same setup as in Theorem 1.*

1. *Vanishing spike: Asymptotic bias is $\frac{1}{T-1}$.*
2. *Constant spike: Let $r = \frac{a}{a+b}$. Asymptotic bias is $\mathbb{E}\left(\frac{(1-r)(rC^2+1-r(T-1))}{rC^2+(1-r)(T-1)}\right)$ where $C^2 \sim \chi^2_{T-1}$.*
3. *Expanding spike: Asymptotic bias is 0.*

Figure 3 shows the asymptotic distribution of $\hat{\tilde{\lambda}}_1$ in the constant spike case for four values of the true $\tilde{\lambda}_1$ and for various T dimensions. The asymptotic distribution explains the behavior of the bias that we observed in the calibrated simulations above. It is clear that the bias is large for low values of the true $\tilde{\lambda}_1$, and tends to decrease as the true $\tilde{\lambda}_1$ increases. This is especially problematic: the bias is greater when there is little shared risk between farmers. This is likely to be the case in places where farmers are more heterogeneous, such as smallholder systems in developing countries. It is also clear that the bias is a function of T: increasing T reduces the bias, and for values as high as $T = 100$, this bias appears negligible. Unfortunately, having 100 years of field-level data ($T = 100$) is unrealistic, and climate change would limit the usefulness of such a long series even if it were available. Looking at a value of $T = 20$ which is more realistic for agricultural data, bias is still present, in particular for lower values of $\tilde{\lambda}_1$. This suggests that even with high-quality data, an insurer will still be over-confident in her assessment of the quality of an insurance product, unless the true quality is very high.

Ideally, deriving the bias would help construct a bias-corrected estimator. Unfortunately, the extent of the bias depends itself on the true yet unknown value of $\tilde{\lambda}_1$. This is a challenging statistical problem, which we leave for further research. However, the result we obtained can be used to derive an upper bound on the bias. Note indeed that the bias is maximum at $\tilde{\lambda}_1 = 0$, taking a value of $1/(T-1)$. This suggests a simple rule of thumb for practitioners to quantify the expected bias they can face in the worst case scenario.

Figure 3: Asymptotic distribution of $\hat{\tilde{\lambda}}_1$



4.3 Illustration of the theorem

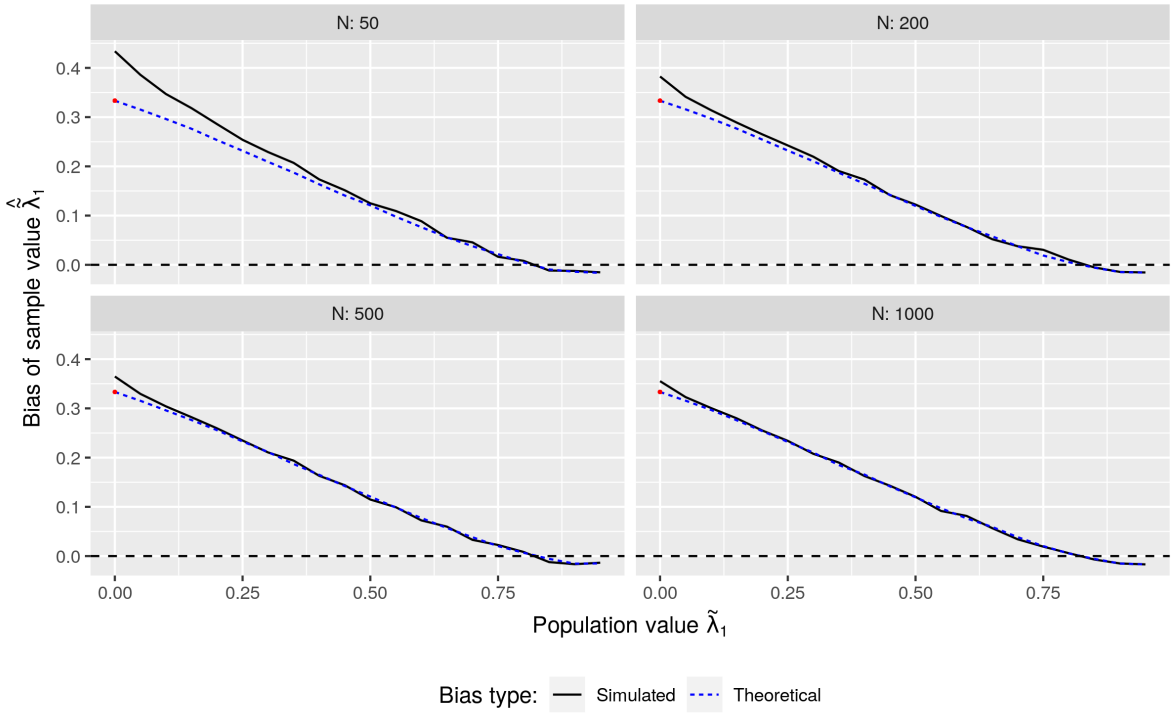
To illustrate this result, we simulate population covariance matrices $\Sigma_N(\tilde{\lambda})$ according to the moderate spike model, varying $\tilde{\lambda}_1$ between 0 and 1. That is, $\Sigma_N(\tilde{\lambda}) = Q_N \Lambda_N(\tilde{\lambda}) Q_N^{'}$, where $\Lambda_N(\tilde{\lambda})$ is a diagonal matrix with the eigenvalues $[aN = \tilde{\lambda}/(1-\tilde{\lambda})N, 1, \ldots, 1]$ and Q_N is a random orthogonal matrix. For each of the $\Sigma(\tilde{\lambda})$ covariances, we then simulate data with sample size of $T \in [4, 20, 100]$ and dimension $N \in [50, 200, 500, 1000]$ assuming a multivariate normal distribution that is i.i.d. over time. The correlation metrics being insensitive to the values of the means vector μ_N, we simply set it to 0. In other words, we now simulate data from:

$$Y_{T,N} \overset{iid}{\sim} \mathcal{N}\left(0_N, \Sigma_N(\tilde{\lambda}) = Q \Lambda_N(\tilde{\lambda}) Q^{'}\right)$$

We then estimate $\hat{\tilde{\lambda}}_1$ on the simulated data. Figure 4 shows the resulting bias estimates on the y-axis, and the true population value $\tilde{\lambda}_1$ on the x-axis. The black line represent the bias of the values estimated on the simulated data from the spiked model, and the blue line represents the formula from Theorem (1). The red dot corresponds to the worst bound $1/(T-1)$. Focusing first on the behaviour of the bias, we clearly see the phenomenon observed with the SCYM-calibrated simulations shown in section 3. For low values of $\tilde{\lambda}$, we observe a very substantive upwards bias, leading to over-confidence in the quality of an insurance product. This bias decreases with increasing $\tilde{\lambda}$, and even reverses at very high values of $\tilde{\lambda}$. Comparing now the difference between the simulated bias (black line) and our asymptotic formula from (1), we see that the formula approximates remarkably well the empirical bias for dimensions as low as $N = 200$. More interestingly, having a higher dimensionality N is no longer a curse, but improves instead the validity of the bias formula!

As a final test, we compare the empirical bias obtained in the calibrated simulations to the analytical bias formula. Remember that the calibrated simulations were generated based on the empirical covariance matrices estimated from the satellite data. As such, there is no guarantee that the asymptotic bias formula

Figure 4: Theoretical bias and bias from the spiked-model simulations



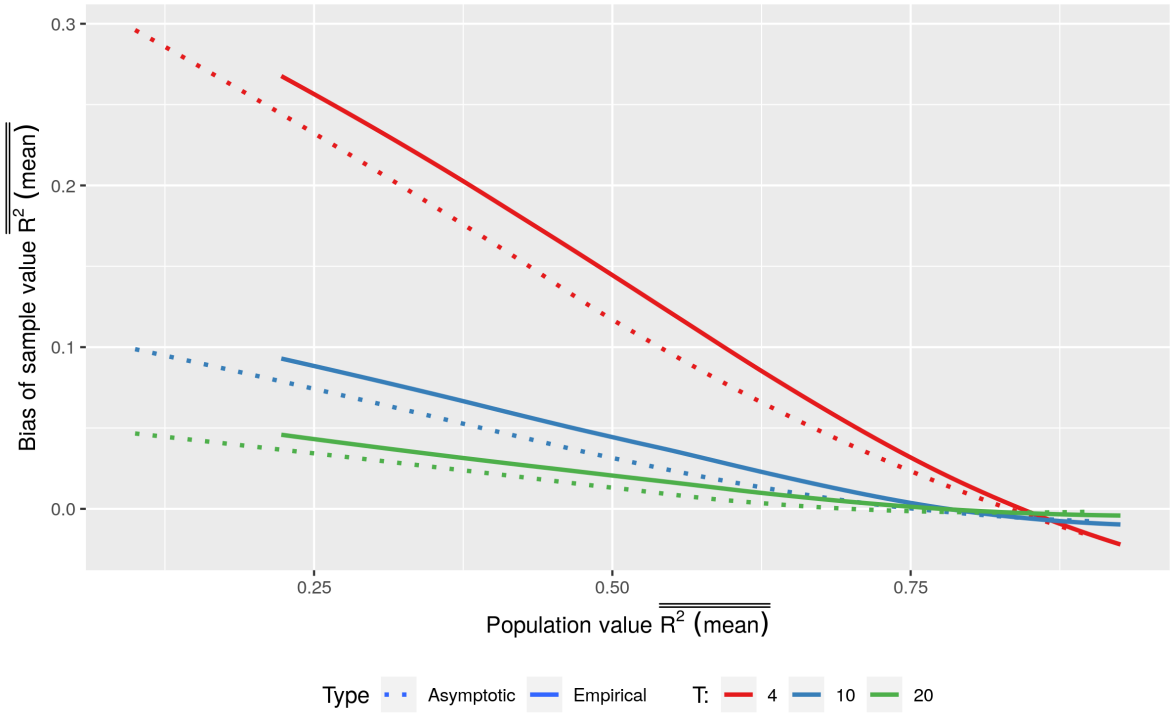
holds, since the latter is obtained assuming a constant spike model, which might not hold in the SCYM dataset. To further test the generality of the bias formula, we focus on the total R^2 measure based on the area-yield index, instead of focusing on $\hat{\tilde{\lambda}}_1 = R^2(w^\star)$. In a constant spike model, one can show that these measures are very similar, but this is not necessarily the case in practice. Figure 5 shows the empirical bias (straight line) together with the bias predicted by theorem 2 (dotted line). The analytical bias formula approximates the empirical bias well, with a difference that tends to vanish as the dimension N increases. This result is very encouraging, suggesting that our choice of a constant spike model to represent the data might be relevant in practice. Furthermore, it indicates that the theory developed here provides a reliable tool for practitioners to assess the potential bias in their measures of basis risk.

5 Conclusion

High resolution satellite images coupled with recent advances in machine learning are expected to yield significant progress in index insurance. However, the discussion to date has overlooked the fact that while data is becoming much richer in terms of number of fields observed, the number of years we observe remains low. As we have shown in this paper both theoretically and empirically through simulation, this introduces a downward bias in common measures of basis risk, which if left uncorrected is likely to yield overly optimistic assessments of insurance product quality.

Our paper is the first to identify this bias in commonly used measures of index insurance quality, which is important because it is likely to lead to overly optimistic assessments of product quality. The theory we develop to explain the observed bias provides a useful set of building blocks to approximate and bound the bias in real-world situations. Academics as well as organizations and governments designing new index insurance products ought to take this into account, especially when products are being designed based on a small number of time periods; the simulation methods we develop provide a strategy for generating more

Figure 5: Theoretical bias and bias from the SCYM-calibrated simulations



accurate product quality estimates.

We highlight two aspects of this bias that are especially problematic for developing country settings where satellite-based index insurance has the greatest potential benefit. First, the bias is generally larger when individual farm yields are less correlated, which is more true in rural smallholder systems than in large developed country farms. Second, the bias can actually increase as the number of fields in the data increases, meaning high resolution data may yield worse estimates of basis risk if the bias is uncorrected. For these reasons, it is critical that companies, governments, and non-governmental organizations designing index insurance products are aware of and take steps to correct the bias we study in this paper.

Future work ought to focus on developing methods to correct the bias identified in this paper. More careful modeling of the covariance between fields is a promising avenue: methods that take into account spatial correlation and/or allow the covariance matrix between fields to vary over time are promising. For example, in the worst years, when poor weather conditions are the dominant factor affecting yields, it may be that field-level outcomes are more correlated than in more typical years, when individual non-weather shocks drive most of the variation.

This paper showed evidence of bias in both linear and quantile-based measures of basis risk, yet confirmed analytically this bias only in the linear case. Focusing on linear measures of basis risk is convenient in that they are easy to understand and study analytically. At the same time, linear measures fail to capture the fact that the effect of basis risk is nonlinear: failures to accurately predict negative shocks are much more detrimental than failures in good years. Analyzing the theoretical bias of nonlinear measures of basis risk and of expected utility metrics remains a challenging task where future research will be needed.

A Appendix

Notation. Let X_n be a sequence of random variables with distribution function F_n respectively and let X be another random variable with distribution function F. Then, we say that X_n converges in distribution to X, and write $X_n \xrightarrow{d} X$ if for all continuity points x of F, $F_n(x) \to F(x)$ as $n \to \infty$. We say that X_n converges in probability to X, and write $X_n \xrightarrow{p} X$ if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$, for any fixed $\epsilon > 0$. Finally, we say that X_n converges almost surely to X, and write $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}(X_n \to X) = 1$. It is well known that $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$. For details, see Billingsley (1995). □

Proof of Theorem 1. Let S be the sample covariance matrix for the data $Y_1, \cdots, Y_T$, i.e. $S = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \overline{Y})(Y_t - \overline{Y})'$ where $\overline{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_i$. Recall that $\hat{\tilde{\lambda}} = \lambda_{max}(S)/Trace(S)$. It is well known from standard multivariate theory (see Mardia et al. (1979)) that $S \overset{d}{=} \frac{1}{T}X'X = \frac{1}{T}\sum_{t=1}^{T-1} X_i X_i'$ where X is a $(T-1) \times N$ matrix with rows $X_i \overset{iid}{\sim} \mathcal{N}_N(0, \Sigma)$. Hence, in whatever follows, we work with the iid data $X_1, \cdots, X_{T-1}$. Also, define the dual sample covariance matrix for the data X, as $S_D = \frac{1}{T-1}XX'$.

For the vanishing spike case, we use Theorem 1 from Ahn et al. (2007), see also Ishii et al. (2014). Note that the function $A \mapsto \lambda_{max}(A)/Trace(A)$ is continuous on the space of positive definite matrices. So, we conclude that $\frac{\lambda_{max}(S_D)}{Trace(S_D)} \xrightarrow{a.s.} \frac{1}{T-1}$ and therefore $\frac{\lambda_{max}(X'X/T)}{Trace(X'X/T)} \xrightarrow{a.s.} \frac{1}{T-1}$. Due to the distributional equivalence between S and $\frac{1}{T}X'X$ we conclude that $\lambda_{max}(S)/Trace(S) \xrightarrow{a.s.} \frac{1}{T-1}$.

For the constant and expanding spike cases, we follow the method presented in Section 4 of Ahn et al. (2007). The dual covariance matrix S_D can be expressed as $(T-1)S_D = aN^{\alpha}Z_1Z_1 + b\sum_{i=2}^{N} Z_iZ_i'$ where Z_i are iid $\mathcal{N}_{T-1}(0, I)$ where I denotes the $(T-1)-$dimensional identity matrix. Hence, $(T-1)S_D/N^{\alpha}$ converges a.s. to $aZ_1Z_1' + bI$ if $\alpha = 1$ (constant spike) and to aZ_1Z_1' if $\alpha > 1$ (expanding spike). Hence, $\lambda_{max}(S_D)/Trace(S_D) \xrightarrow{d} \frac{aC^2+b}{aC^2+b(T-1)}$ in the constant spike case and in probability to 1 in the expanding spike case. Here, $C^2 = \|Z_1\|^2 \sim \chi^2_{T-1}$. Again, utilizing the distributional equivalence between S and $X'X/T$ we conclude that $\lambda_{max}(S)/Trace(S) \xrightarrow{d} \frac{aC^2+b}{aC^2+b(T-1)}$ in the moderate spike case, and $\lambda_{max}(S)/Trace(S) \xrightarrow{p} 1$ in the expanding spike case. This completes the proof.

□

Proof of Theorem 2. Note that the random variables $\hat{\tilde{\lambda}}$ are bounded between 0 and 1. By Theorem 1, in each of vanishing, constant and expanding spike cases, this sequence of random variables converges in distribution, and hence by uniform integrability, their expectations also converge to the corresponding quantities. This shows that for the vanishing and expanding spike cases, $\mathbb{E}(\lambda_{max}(S)/Trace(S))$ converges to $\frac{1}{T-1}$ and 1 respectively. Noting that the true asymptotic largest eigenvalue shares were 0 and 1 respectively in the vanishing and expanding spike cases, the biases are $1/(T-1)$ and 0 respectively.

For the constant spike case, writing $r = a/(a+b)$, we get that

$$Bias(\hat{\tilde{\lambda}}) = \mathbb{E}(\hat{\tilde{\lambda}}) - r \to \mathbb{E}\left(\frac{rC^2 + (1-r)}{rC^2 + (1-r)(T-1)}\right) - r = \mathbb{E}\left(\frac{(1-r)(rC^2 + 1 - r(T-1))}{rC^2 + (1-r)(T-1)}\right)$$

□

References

Ahn, J., J. S. Marron, K. M. Muller, and Y.-Y. Chi (2007): "The High-Dimension, Low-Sample-Size Geometric Representation Holds under Mild Conditions," *Biometrika*, 94, 760–766.

Aoshima, M., D. Shen, H. Shen, K. Yata, Y.-H. Zhou, and J. S. Marron (2018): "A survey of high dimension low sample size asymptotics," *Australian & New Zealand Journal of Statistics*, 60, 4–19.

Azzari, G. and D. Lobell (2017): "Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring," *Remote Sensing of Environment*, 202, 64 – 74, big Remotely Sensed Data: tools, applications and experiences.

Benami, E., Z. Jin, M. R. Carter, A. Ghosh, R. J. Hijmans, A. Hobbs, B. Kenduiywo, and D. B. Lobell (2021): "Uniting Remote Sensing, Crop Modeling, and Economics for Agricultural Risk Management," *Nature Reviews Earth and Environment*, 21:2, 140–159.

Billingsley, P. (1995): *Probability and Measure, 3 edition*, Wiley Series in Probability and Mathematical Statistics.

Bokusheva, R. (2018): "Using copulas for rating weather index insurancecontracts, Journal of Applied Statistics," *Journal of Applied Statistics*, 45:13, 2328–2356.

Bucheli, J., T. Dalhaus, and R. Finger (2020): "The optimal drought index for designing weather index insurance," *European Review of Agricultural Economics*, jbaa014.

Burke, M. and D. B. Lobell (2016): "Satellite-based assessment of yield variation and its determinants in smallholder African systems," Tech. rep., Department of Earth System Science, Stanford University.

Carter, M., A. de Janvry, E. Sadoulet, and A. Sarris (2017): "Index insurance for developing country agriculture: a reassessment," *Annual Review of Resource Economics*, 9, 421–438.

Carter, M., A. De Janvry, E. Sadoulet, A. Sarris, et al. (2014): "Index-based weather insurance for developing countries: A review of evidence and a set of propositions for up-scaling," *Development Policies working paper*, 111.

Clarke, D. J. (2016): "A Theory of Rational Demand for Index Insurance," *American Economic Journal: Microeconomics*, 8, 283–306.

Clarke, D. J., D. Clarke, O. Mahul, K. N. Rao, and N. Verma (2012): "Weather based crop insurance in India," *World Bank Policy Research Working Paper*.

Conradt, S., R. Finger, and R. Bokusheva (2015): "Tailored to the extremes: Quantile regression for index-based insurance contract design," *Agricultural Economics*, 46, 537–547.

De Leeuw, J., A. Vrieling, A. Shee, C. Atzberger, K. M. Hadgu, C. M. Biradar, H. Keah, and C. Turvey (2014): "The Potential and Uptake of Remote Sensing in Insurance: A Review," *Remote Sensing*, 6, 10888–10912.

Deines, J. M., R. Patel, S.-Z. Liang, W. Dado, and D. B. Lobell (2021): "A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt," *Remote Sensing of Environment*, 253, 112174.

Deines, J. M., S. Wang, and D. B. Lobell (2019): "Satellites reveal a small positive yield effect from conservation tillage across the US Corn Belt," *Environmental Research Letters*, 14, 124038.

Elabed, G., M. F. Bellemare, M. R. Carter, and C. Guirkinger (2013): "Managing basis risk with multiscale index insurance," *Agricultural Economics*, 44, 419–431.

Hall, P., J. S. Marron, and A. Neeman (2005): "Geometric Representation of High Dimension, Low Sample Size Data," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 427–444.

Ishii, A., K. Yata, and M. Aoshima (2014): "Asymptotic Distribution of the Largest Eigenvalue via Geometric Representations of High-Dimension, Low-Sample-Size Data," *Sri Lankan Journal of Applied Statistics*, 5(4), 81–94.

Jensen, N. D., C. B. Barrett, and A. G. Mude (2016): "Index Insurance Quality and Basis Risk: Evidence from Northern Kenya," *American Journal of Agricultural Economics*, 98, 1450–1469.

Jin, Z., G. Azzari, M. Burke, S. Aston, and L. David (2017a): "Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa." *Remote Sensing*, 9(9):931.

Jin, Z., G. Azzari, and D. B. Lobell (2017b): "Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches," *Agricultural and Forest Meteorology*, 247, 207 – 220.

Jin, Z., G. Azzari, C. You, S. Di Tommaso, S. Aston, M. Burke, and D. B. Lobell (2019): "Smallholder maize area and yield mapping at national scales with Google Earth Engine," *Remote Sensing of Environment*, 228, 115 – 128.

Johnstone, I. M. (2001): "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *The Annals of Statistics*, 29, 295–327.

Johnstone, I. M. and D. Paul (2018): "PCA in High Dimensions: An Orientation," *Proceedings of the IEEE*, 106, 1277–1292.

Jung, S. and J. S. Marron (2009): "PCA consistency in high dimension, low sample size context," *Ann. Statist.*, 37, 4104–4130.

Koenker, R. and J. A. F. Machado (1999): "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Lobell, D. B., G. Azzari, M. Burke, S. Gourlay, Z. Jin, T. Kilic, and S. Murray (2020): "Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis," *American Journal of Agricultural Economics*, 102, 202–219.

Lobell, D. B., D. Thau, C. Seifert, E. Engle, and B. Little (2015): "A scalable satellite-based crop yield mapper," *Remote Sensing of Environment*, 164, 324 – 333.

Mahul, O. (1999): "Optimum Area Yield Crop Insurance," *American Journal of Agricultural Economics*, 81, 75–82.

Mardia, K., J. Kent, and J. Bibby (1979): *Multivariate Analysis, 1 edition*, Academic Press.

Miranda, M. J. (1991): "Area-Yield Crop Insurance Reconsidered," *American Journal of Agricultural Economics*, 73, 233–242.

Seifert, C. A., G. Azzari, and D. B. Lobell (2018): "Satellite detection of cover crops and their effects on crop yield in the Midwestern United States," *Environmental Research Letters*, 13, 064033.

Stigler, M. (2018): "Supply response at the field-level: disentangling area and yield effects," Tech. rep., UC Davis, ARE.

Stigler, M. and D. Lobell (2020): "Suitability of index insurance: new insights from satellite data," Tech. rep., Stanford University.

——— (2021): "Optimal index insurance and basis risk decomposition: an application to Kenya," Tech. rep., Stanford University.

Wang, W. and J. Fan (2017): "Asymptotics of empirical eigenstructure for high dimensional spiked covariance," *Ann. Statist.*, 45, 1342–1374.