ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Design-free estimation of variance matrices

Karim M. Abadir [a], Walter Distaso [a,*], Filip Žikeš [b]

[a] Imperial College London, United Kingdom

[b] Bank of England, United Kingdom

ARTICLE INFO

Article history:
Received 11 March 2013
Received in revised form
19 March 2014
Accepted 27 March 2014
Available online xxxx

Keywords:
Variance matrices
Ill-conditioning
Mean squared error
Mean absolute deviations
Resampling
U-statistics

ABSTRACT

This paper introduces a new method for estimating variance matrices. Starting from the orthogonal decomposition of the sample variance matrix, we exploit the fact that orthogonal matrices are never ill-conditioned and therefore focus on improving the estimation of the eigenvalues. We estimate the eigenvectors from just a fraction of the data, then use them to transform the data into approximately orthogonal series that deliver a well-conditioned estimator (by construction), even when there are fewer observations than dimensions. We also show that our estimator has lower error norms than the traditional one. Our estimator is design-free: we make no assumptions on the distribution of the random sample or on any parametric structure the variance matrix may have. Simulations confirm our theoretical results and they also show that our simple estimator does very well in comparison with other existing methods.

1. Introduction

Apart from calculating the mean, estimating the variance of a random vector is the most basic problem in statistics. It has numerous applications in sciences, social sciences, and humanities. Examples go from financial time series, where variance matrices are used as a measure of risk, to molecular biology, where they are used for gene classification purposes. Yet the estimation of variance matrices is a statistically challenging problem, since the number of parameters grows as a quadratic function of the number of variables. To make things harder, conventional methods deliver nearly-singular (ill-conditioned) estimators when the dimension k of the matrix is large relative to the sample size n. As a result, estimators are very imprecise and operations such as matrix inversions amplify the estimation error further.

One strand of the literature has tackled this problem by trying to come up with methods that are able to achieve a dimensionality reduction by exploiting sparsity, imposing zero restrictions on some elements of the variance matrix. Wu and Pourahmadi (2003) and Bickel and Levina (2008a) propose banding methods to find consistent estimators of variance matrices and their inverse. Other authors resort to thresholding (Bickel and Levina, 2008b; El Karoui, 2008; Fan et al., forthcoming) or penalized likelihood methods (see, e.g., Fan and Peng, 2004 for the underlying general theory) to estimate sparse large variance matrices. Notable examples of papers using the latter method are Huang et al. (2006), Rothman et al. (2008, 2009). Recently, Lam and Fan (2009) proposed a unified theory of estimation, introducing the concept of *sparsistency*, which means that (asymptotically) the zero elements in the matrix are estimated as zero almost surely.

An alternative approach followed by the literature is to achieve dimensionality reduction using factor models. The idea is to replace the k individual series with a small number of unobservable factors such that they are able to capture most of the variation contained in the original data. Interesting examples are given by Fan et al. (2008), Wang et al. (2009) and Lam and Yao (2012). Fan et al. (2011) combine a factor structure with sparsity of the variance matrix.

A third route is given by shrinkage, which entails substituting the original ill-conditioned estimator with a convex combination including it and a target matrix. The original idea is due to Stein (1956), where it was applied to the estimation of the mean vector. Applications to variance matrix estimation include Jorion (1986), Muirhead (1987) and Ledoit and Wolf (2003, 2004a,b, 2012). Intuitively, the role of the shrinkage parameter is to balance the estimation error coming from the ill-conditioned variance matrix and the specification error associated with the target matrix. Ledoit and Wolf (2003) propose an optimal estimation procedure for the shrinkage parameter, where the chosen metric is the Frobenius norm between the variance and the shrinkage matrix. An alternative approach whereby off-diagonal elements are downweighted towards zero is given in McMurry and Politis (2010) and Politis

* Correspondence to: Imperial College Business School, Imperial College London, Exhibition Road, SW7 2AZ, London, United Kingdom. Tel.: +44 207 594 3293; fax: +44 207 594 9189.

E-mail address: w.distaso@imperial.ac.uk (W. Distaso).

http://dx.doi.org/10.1016/j.jeconom.2014.03.010
0304-4076/

(2011) in the context of time series. See also an approach to shrinkage via condition-number regularization in Won et al. (2013).

In this paper, we introduce a new method to estimate nonsingular variance matrices. We propose a different approach for tackling this problem. Starting from the orthogonal decompositions of symmetric matrices, we exploit the fact that orthogonal matrices are never ill-conditioned (they have the perfect condition number of 1), thus identifying the source of the problem as the eigenvalues. Our task is then to come up with an improved estimator of the eigenvalues. We achieve this by estimating the eigenvectors from just a fraction of the data (a subsample), then using them to transform the data into approximately orthogonal series that we use to estimate a well-conditioned matrix of eigenvalues. Effectively, this simple idea reduces the multivariate problem to k univariate ones that are easy to solve. Moreover, we improve precision further by repeating our procedure over different subsamples, and we show that averaging the resulting estimators leads to a superior performance.

Even though we only use the simple traditional formula for the sample variance matrix in both steps of our basic orthogonalization-estimation procedure, the result is a well-conditioned and precise estimator. Because of the orthogonalization of the data, the resulting estimate is positive definite with probability one, *even* when the dimension of the matrix is larger than the sample size: $k > n$. Our estimator outperforms the traditional one, not only by achieving a substantial improvement in the condition number, but also by large improvements in error norms that measure its deviation from the true variance matrix. We also show that our simple estimator does very well in comparison to other existing methods.

Our method has a number of other attractive features. First, it is design-free, in the sense that no assumptions are made on the densities of the random sample or on any underlying parametric model for the structure of the variance matrix. Second, it always delivers nonsingular well-conditioned estimators, hence remaining precise when further operations (such as inversions) are required. Such operations are trivially easy to implement in our setup, since matrix functions are efficiently written in terms of eigenvalues and eigenvectors; e.g., see Abadir and Magnus (2005, Ch. 9).

This paper is organized as follows. Section 2 introduces the proposed estimator in its simplest baseline then general versions, and establishes its main properties. Section 3 studies in a Monte-Carlo experiment the finite-sample properties of our estimator and how it compares with other methods. It also provides further guidance on its use in practice. Section 4 concludes. The derivations are collected in the Appendix.

2. The new estimator

This section contains two parts. First, we briefly present the setup and the intuition for why the new estimator will perform well. Second, we investigate the properties for the simplest baseline formulation of our estimator, and afterwards we tackle the full version of it as an extension for which the properties are then easily obtained. We describe the optimal choice of two subsampling parameters (one for each step of the baseline orthogonalization-estimation procedure), first in the case of fixed k, then when k expands as n increases.

2.1. The setup and the main idea behind the orthogonalization-estimation procedure

Let $\boldsymbol{\Sigma} := \text{var}(\boldsymbol{x})$ be a finite $k\times k$ positive definite variance matrix of $\boldsymbol{x}$. Suppose we have an i.i.d. sample $\{\boldsymbol{x}_i\}_{i=1}^n$, arranged into the $n\times k$ matrix $\boldsymbol{X} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ on which we base the usual estimator (ill-conditioned when k is large relative to n)

$$\widehat{\boldsymbol{\Sigma}} \equiv \widehat{\text{var}}(\boldsymbol{x}) := \frac{1}{n}\boldsymbol{X}'\boldsymbol{M}_n\boldsymbol{X},$$

where $\boldsymbol{M}_n := \boldsymbol{I}_n - \frac{1}{n}\iota_n\iota_n'$ is the de-meaning matrix of dimension n and ι_n is a $n\times 1$ vector of ones. The assumption of an i.i.d. setup is not as restrictive as it may seem: the data can be filtered by an appropriate model (rather than just de-meaning by $\boldsymbol{M}_n$) and the method applied to the residuals; for example, fitting a VAR model (if adequate) to a vector of time series and applying the method to the residuals. We will stick to the simplest setup, so as to clarify the workings of our method.

We can decompose this symmetric matrix as

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{P}}', \tag{1}$$

where $\widehat{\boldsymbol{P}}$ is orthogonal and has typical column $\widehat{\boldsymbol{p}}_i$ ($i = 1, \ldots, k$), $\widehat{\boldsymbol{\Lambda}}$ being the diagonal matrix of eigenvalues of $\widehat{\boldsymbol{\Sigma}}$. The condition number of any matrix is the ratio of the largest to smallest singular values of this matrix, a ratio of 1 being the lowest (best numerical) condition. By orthogonality, all the eigenvalues of $\widehat{\boldsymbol{P}}$ lie on the unit circle and this matrix is always well-conditioned for any n and k. This leaves $\widehat{\boldsymbol{\Lambda}}$ as the source of the ill-conditioning of the estimate $\widehat{\boldsymbol{\Sigma}}$. We will therefore consider an improved estimator of $\boldsymbol{\Lambda}$: a simple estimator of $\boldsymbol{P}$ will be used to transform the data to achieve approximate orthogonality of the transformed data (in variance terms), hence yielding a better-conditioned estimator of the variance matrix.

We can rewrite the decomposition (1) as

$$\widehat{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{P}}'\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{P}} = \text{diag}(\widehat{\text{var}}(\widehat{\boldsymbol{p}}_1'\boldsymbol{x}), \ldots, \widehat{\text{var}}(\widehat{\boldsymbol{p}}_k'\boldsymbol{x})) \tag{2}$$

the last equality following since $\widehat{\boldsymbol{\Lambda}}$ is diagonal by definition. Now suppose that, instead of basing $\widehat{\boldsymbol{P}}$ on the whole sample, we base it on only m observations (say the first m ones, since the i.i.d. setup means that there is no gain from doing otherwise), use it to approximately orthogonalize the rest of the $n-m$ observations (as $\widehat{\boldsymbol{p}}_i'\boldsymbol{x}$ did in (2) for all the observations) which are then used to reestimate $\boldsymbol{\Lambda}$. Taking $m\to\infty$ and $n-m\to\infty$ as $n\to\infty$, standard statistical analysis implies that the resulting estimators are consistent. Notice that the choice of basing the second step on the remaining $n-m$ observations comes from two considerations. First, it is inefficient to discard observations in an i.i.d. setup, so we should not have fewer than these $n-m$ observations. Second, we should not reuse some of the first m observations because they worsen the estimate of $\boldsymbol{\Lambda}$: this will be seen in Proposition 2 (for the condition number) and implied by the estimators' expansions in Propositions 3–4 (for the error norms). As a result, m becomes the only remaining subsampling parameter in question. Proposition 3–4 will show that the precision of the new estimator is optimized by expressing m as a function of n asymptotically. Proposition 5 then extend these results to the case when k varies as n increases, and only then do we consider the alternative definition of consistency as convergence in mean square. These propositions are followed by a concluding discussion of how to calculate the optimal m by resampling in any finite sample, not just asymptotically.

Intuitively, by orthogonalizing the data, our estimator reduces the multivariate problem of ill-conditioning and imprecision to a univariate one for each of the diagonal elements of (2), for which there is a simple positive definite solution even by traditional methods of estimation. The result is a well-conditioned estimator of $\boldsymbol{\Sigma}$, even when $k\geq n$ and the traditional $\widehat{\boldsymbol{\Sigma}}$ is not positive definite. We will prove this in the next subsection.

Another advantage of our procedure is that we can estimate the matrix itself as well as any function thereof in one go from the eigenvalue decomposition. The other methods seen in the introduction focus on the variance matrix, and if a function is needed (such as the inverse), one has to make further multivariate calculations to obtain it. This can be imprecise if the dimension is large. In addition to the advantages seen so far, we will show that also the precision of our estimator is an advantage, even though we only use the simple traditional sample variance estimator in both steps of our procedure.

Please cite this article in press as: Abadir, K.M., et al., Design-free estimation of variance matrices. Journal of Econometrics (2014), http://dx.doi.org/10.1016/j.jeconom.2014.03.010