Biometrika (2025), **112**, 4, asaf052 https://doi.org/10.1093/biomet/asaf052
Advance Access publication 21 July 2025

Dimension estimation in a spiked covariance model using high-dimensional data augmentation

By U. RADOJIČIĆ

Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedener Hauptstrasse 8–10, Vienna 1040, Austria
una.radojicic@tuwien.ac.at

and J. VIRTA

Department of Mathematics and Statistics, University of Turku, Turku 20014, Finland
joni.virta@utu.fi

Summary

We propose a modified, high-dimensional version of a recent dimension estimation procedure that determines the dimension via the introduction of augmented noise variables into the data. Our asymptotic results show that the proposal is consistent in wide, high-dimensional scenarios, and further shed light on why the original method breaks down when the dimension of either the data or the augmentation becomes too large. Simulations and real data are used to demonstrate the superiority of the proposal to competitors both under and outside of the theoretical model.

Some key words: Augmentation; Covariance matrix; Low-rank model; Order determination.

1. Introduction

We revisit the problem of estimating the number of spikes in a spiked covariance model; see, e.g., Luo & Li (2016), Nordhausen et al. (2022) and Bernard & Verdebout (2024). We assume that $x_{1,n}, \ldots, x_{n,n}$ is a random n-indexed sample (a triangular array) drawn from the p_n-variate normal distribution $\mathcal{N}_{p_n}(\mu_n, \Sigma_n)$, where the covariance matrix Σ_n has eigenvalues $\lambda_1 + \sigma^2 > \cdots > \lambda_d + \sigma^2 > \sigma^2 = \cdots = \sigma^2$. The assumption of distinct spikes is for mathematical convenience. The constants $\lambda_1, \ldots, \lambda_d, \sigma^2$ and the true signal dimension $d \geqslant 1$ do not vary with n, and increasing n simply has the effect of adding more noise dimensions in the model. Such models are known as spiked covariance structures (Yao et al., 2015). While consistent estimation of μ_n and Σ_n typically requires structural assumptions, such as sparsity, when $p_n/n \not\to 0$, our proposed estimator of d does not rely on such assumptions. Thus, our objective is to estimate the fixed dimension d from the sample $x_{1,n}, \ldots, x_{n,n}$, without imposing additional strong assumptions on μ_n and Σ_n.

Most existing estimators for d rely on subsphericity testing, information-theoretic criteria or risk minimization. A different approach is predictor augmentation (PA), introduced by Luo & Li (2021). However, its theoretical justification hinges on the assumption that the sample covariance matrix is a consistent estimator of its population counterpart; an assumption that breaks down in high-dimensional regimes where the number of variables grows proportionally with the sample size. This renders the consistency guarantees of PA inapplicable in high-dimensional settings. A full description of the method is given in § 2, but on a heuristic level, in PA the observed $n \times p$ data are augmented into

a sample of size $n \times (p+r)$, where the added nr variables are independent and identically distributed Gaussian with a specific variance. Heuristically, the added variables, being pure noise, get mixed with the actual noise subspace, allowing one to pinpoint the jump from the signal subspace to the noise subspace in the spectrum of the covariance matrix of the augmented data. The outcome of predictor augmentation is a function $\phi_n\colon \{0,\dots,p\} \to \mathbb{R}$, whose minimizer d_n is the estimate of the latent dimension d. Luo & Li (2021, Theorem 5) showed that this estimator is, for every fixed r, consistent as $n \to \infty$ under certain mild conditions. However, this result assumes finite p and it turns out that the consistency can fail in high-dimensional scenarios where the dimension $p \equiv p_n$ and/or the number of augmentations $r \equiv r_n$ grow with n for the reasons given above.

Figure 1 illustrates this. It shows the mean augmentation curves ϕ_n (points) over 500 replicates when $n = 400$, $r_n \in \{0.01n, 0.5n\}$, $p_n \in \{0.025n, 0.25n\}$. The true dimension is $d = 1$ with signal-to-noise ratio 2, so ideally $\phi_n(k)$ is minimized at $k = 1$. The lines show the limits of $\phi_n(k)$ derived in § 3. The results reveal that PA yields the correct estimate only when both p_n and r_n are small relative to n; see Luo & Li (2021, Theorem 4). Notably, even when p_n is small ($0.025n$), the method fails if r_n is too large, going against the suggestion of Radojičić et al. (2025), who extended PA to tensors and argued that larger r are beneficial for fixed p. As implied earlier, the failure of PA is due to the fact that the high-dimensional consistency result (Luo & Li, 2021, Theorem 4) would require the sample covariance matrix to be consistent, an assumption invalidated as soon as p_n grows proportional to n; see Fan et al. (2008, Theorem 1). As such, the purpose of the current work is three-fold. (i) We carefully investigate the inconsistency of the augmentation estimator by studying its asymptotics in the doubly high-dimensional setting where both the data dimension p_n and the augmentation dimension r_n diverge to infinity such that $(p_n/n, r_n/n) \to (\gamma_p, \gamma_r) > 0$. In particular, we show that, for every γ_p, there exist rates γ_r, which lead to inconsistent predictor augmentation. (ii) We derive a corrected estimator that is consistent in high-dimensional settings under very mild conditions. We emphasize that, after having estimated the rank, additional assumptions can be imposed on the parameters to estimate them with a chosen method. For example, sparse PCA can consistently estimate the signal subspace under sparsity (Johnstone & Lu, 2009), but as our focus is solely on the dimension d, and we show that our method is consistent without such conditions, we do not impose them here. (iii) We use simulations and real data to compare our estimator to both the original predictor augmentation and the high-dimensional subsphericity-based estimator of Schott (2006). The results show that our proposal surpasses both competitors, tolerates model deviations exceedingly well and outperforms the original predictor augmentation both theoretically, by maintaining consistency in wide high-dimensional settings, and empirically, as demonstrated in a data example and simulations.

2. PREDICTOR AUGMENTATION ESTIMATOR

To estimate d with predictor augmentation (Luo & Li, 2021), the observations $x_{1,n},\dots,x_{n,n}$ are augmented with simulated noise, i.e., we form the augmented sample $z_{i,n} = (x_{i,n}^{\mathrm{T}}, \sigma_n s_{i,n}^{\mathrm{T}})^{\mathrm{T}}$, $i = 1,\dots,n$, where $s_n \sim \mathcal{N}_{r_n}(0, I_{r_n})$, and σ_n is an estimator of the noise standard deviation σ. The number r_n of augmentations is a tuning parameter. Denoting by S_n the sample covariance matrix of the (p_n+r_n)-dimensional augmented sample, we compute a set $u_{1,n},\dots,u_{p_n,n}$ of any of its leading p_n eigenvectors. Decomposing these as $u_{j,n} = (u_{j,n,A}^{\mathrm{T}}, u_{j,n,B}^{\mathrm{T}}, u_{j,n,C}^{\mathrm{T}})^{\mathrm{T}}$, where the dimensions of the three parts are $d, p_n - d, r_n$, respectively, the augmentation estimator of d is based on the sequence of squared norms $\|u_{1,n,C}\|^2,\dots,\|u_{p,n,C}\|^2$. Heuristically, a jump is seen in the magnitudes of these norms when we cross from the signal subspace into the noise subspace. Luo & Li (2021) further combined the eigenvectors with a standardized scree plot computed from the $p_n + 1$ first eigenvalues $\tau_{1,n},\dots,\tau_{p_n+1,n}$ of S_n, and defined $\phi_n\colon \{0,\dots,p\} \to \mathbb{R}$, acting as

$$\phi_n(k) = \sum_{j=0}^{k} \|u_{j,n,C}\|^2 + \frac{\tau_{k+1,n}}{1 + \sum_{j=1}^{k+1} \tau_{j,n}}, \tag{1}$$

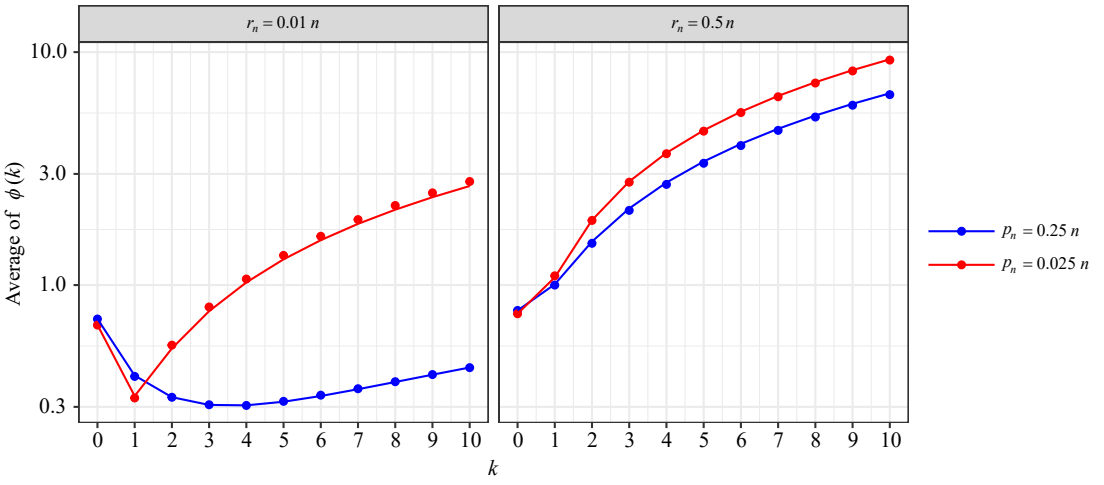

Fig. 1. The average augmentation curves ϕ_n over 500 replicates. The solid lines indicate the pointwise limits.

taking $\|u_{0,n,C}\|^2 = 0$. The estimator of d is then obtained as the minimizer of ϕ_n. Luo & Li (2021) considered independently repeating the augmentation several times and averaging the norms over these. We ignore this possibility, as it does not affect the limiting values of the studied quantities.

In the following sections, we study the behaviour of the above estimator in the asymptotic scenario where $p_n/n \to \gamma_p \in (0, \infty)$ and $r_n/n \to \gamma_r \in (0, \infty)$, as $n \to \infty$, for some constants γ_p, γ_r. Throughout the work, we make the following assumption.

Assumption 1. The smallest signal eigenvalue satisfies $\lambda_d > \sigma^2(\gamma_p + \gamma_r)^{1/2}$.

Assumption 1 is connected to the so-called Baik–Ben Arous–Péché transition that states that the sample estimate of any signal eigenvalue in the interval $(\sigma^2, \sigma^2\{1 + (\gamma_p + \gamma_r)^{1/2}\})$ asymptotically behaves as if it was a noise eigenvalue, with the behaviour also extending to the corresponding eigenvector (Yao et al., 2015). Hence, Assumption 1 is natural and essentially necessary, as without it the effective dimension of our model would actually be smaller than d. Similar assumptions are commonly made; see, e.g., Alaoui et al. (2020), Jagannath et al. (2020) and Mukherjee (2025).

3. Asymptotics of predictor augmentation

For mathematical convenience, we make the simplifying assumption that the noise variance σ^2 is known. The simulations in § 6 below show that the practical behaviour of the method changes remarkably little when σ is replaced with a consistent estimator σ_n of it.

We first establish the limits of the norms $\|u_{j,n,C}\|^2$ of the augmented parts of the eigenvectors.

Theorem 1. *Fix a constant $K \in \mathbb{N}$ such that $K > d$. Under Assumption 1, we have, as $n \to \infty$,*

$$\|u_{j,n,C}\|^2 \xrightarrow{\text{p}} \frac{\gamma_r \sigma^2}{\lambda_j}\left(\frac{\lambda_j + \sigma^2}{\lambda_j + (\gamma_p + \gamma_r)\sigma^2}\right), \qquad j = 1, \ldots, d,$$

$$\|u_{j,n,C}\|^2 \xrightarrow{\text{p}} \frac{\gamma_r}{\gamma_p + \gamma_r}, \qquad j = d+1, \ldots, K.$$

Remark 1. Theorem 1 could still be generalized to have $\min\{p_n + r_n, n-1\}$ in place of the constant K, assuming that $\gamma_p + \gamma_r \neq 1$; see Bloemendal et al. (2016, Theorem 2.17) for details.

As pointed out by a reviewer, the probability limits of $\|u_{1,n,C}\|^2, \|u_{2,n,C}\|^2, \ldots$ form an increasing sequence, with a strict jump between $\|u_{d,n,C}\|^2$ and $\|u_{d+1,n,C}\|^2$. Thus, one way to determine dimension

d would be to use a scree plot of these norms. However, in general, including both eigenvector and eigenvalue information leads to more accurate estimators (see the Supplementary Material for further details), and hence Luo & Li (2021) still combined these norms with the corresponding eigenvalues that they estimated as the largest p_n+1 eigenvalues $\tau_{j,n}$, $j = 1,\dots,p_n+1$, of the sample covariance matrix of the augmented sample $z_{i,n}$. We next give the probability limits of eigenvalues $\tau_{j,n}$, further illustrating the bias of the traditional estimators. Here F_γ^{-1} denotes the quantile function of the Marchenko–Pastur distribution with the concentration parameter $\gamma > 0$.

THEOREM 2. *Under Assumption 1, for the signal eigenvalues, we have*

$$\tau_{j,n} \xrightarrow{\text{p}} (\lambda_j+\sigma^2)\{1+(\gamma_p+\gamma_r)\sigma^2/\lambda_j\}, \qquad j=1,\dots,d,$$

whereas, for the noise eigenvalues, we have the following two cases.

(i) *If $\gamma_p+\gamma_r \in (0,1]$ then, for a sequence $j_n > d$, such that $j_n/(p_n+r_n) \to 1-q \in [0,1]$ as $n\to\infty$, we have $\tau_{j_n,n} \xrightarrow{p} \sigma^2 F^{-1}_{\gamma_p+\gamma_r}(q)$.*

(ii) *If $\gamma_p+\gamma_r \in (1,\infty)$ then, for a sequence $j_n \in [d+1, n-1]$, such that $j_n/(n-1) \to 1-q \in [0,1]$ as $n\to\infty$, we have $\tau_{j_n,n} \xrightarrow{p} \sigma^2(\gamma_p+\gamma_r)F^{-1}_{(\gamma_p+\gamma_r)^{-1}}(q)$, whereas, $\tau_{j,n}=0$ for all $j \geqslant n$.*

Remark 2. Assume that, contrary to Assumption 1, we have a very weak signal, in the sense that $\lambda_d = \sigma^2(\gamma_p+\gamma_r)^{1/2}$. In this case, arguing as in the proofs of Theorems 1 and 2, one can show that both $\|u_{d,n,C}\|^2$ and $\|u_{d+1,n,C}\|^2$ converge to the same constant, as do $\tau_{d,n}$ and $\tau_{d+1,n}$. This further demonstrates that, without Assumption 1, no method based on the asymptotic limits of the eigenvalues and eigenvectors of the sample covariance matrix is able to estimate d.

We denote the probability limits of $\|u_{j,n,C}\|^2$ and $\tau_{j,n}$ by $\|u_{j,C}\|^2$ and τ_j, respectively. Furthermore, the pointwise limit of the objective function ϕ_n in (1) is denoted by ϕ, i.e., $\phi_n(k) \xrightarrow{\text{p}} \phi(k)$.

4. WHEN CAN PREDICTOR AUGMENTATION FAIL?

Theorems 1 and 2 can be used to determine whether any given scenario leads to a consistent augmentation estimate of d. For example, plugging in the simulation scenario in § 1 reveals that, indeed, $\phi(1) > \phi(2)$ holds when $\gamma_p = 0.25$ and $\gamma_r = 0.01$, leading to the inconsistent estimate. Because of the form of the objective function (1), necessary and sufficient conditions for its consistency are not feasible to obtain, but Theorem 3 below lists three 'natural' regions of the parameter space $(\lambda_d, \sigma^2, \gamma_p, \gamma_r)$ where d is not the minimizer of ϕ, i.e., an inconsistency occurs.

THEOREM 3. *Under Assumption 1, the following statements hold.*

(i) *Given any $\lambda_d > 0$ and $\gamma_p > 0$, there exists $\gamma_r^0 \in (0, \lambda_d^2\sigma^{-4}-\gamma_p)$ such that $\phi(d) > \phi(d+1)$ for every $\gamma_r \in (0,\gamma_r^0)$.*

(ii) *Given any $\gamma_p > 0$, there exists small enough $\lambda_d > \sigma^2\sqrt{\gamma_p}$ such that, for all $\gamma_r \in (0, \lambda_d^2\sigma^{-4}-\gamma_p)$, we have $\phi(d) > \phi(d+1)$.*

(iii) *Given any $\lambda_d > 0$, there exists large enough $\gamma_p \in (0, \lambda_d^2\sigma^{-4})$ such that, for all $\gamma_r \in (0, \lambda_d^2\sigma^{-4}-\gamma_p)$, we have $\phi(d) > \phi(d+1)$.*

The upper bounds $\gamma_r < \lambda_d^2\sigma^{-4}-\gamma_p$ in Theorem 3 ensure that γ_r satisfies Assumption 1. Hence, these bounds cannot be improved, as doing so would make the signal dimension unidentifiable. Statement (i) of Theorem 3 shows that there always exists small enough γ_r for which the estimator is inconsistent. Thus, r_n should be chosen large enough relative to p_n and λ_d. In particular, in high-dimensional settings, a finite number of augmentations is always insufficient for consistency. Statement (ii) states that, for any *data-collection* rate γ_p, there always exists a scenario where the signal is strong enough to

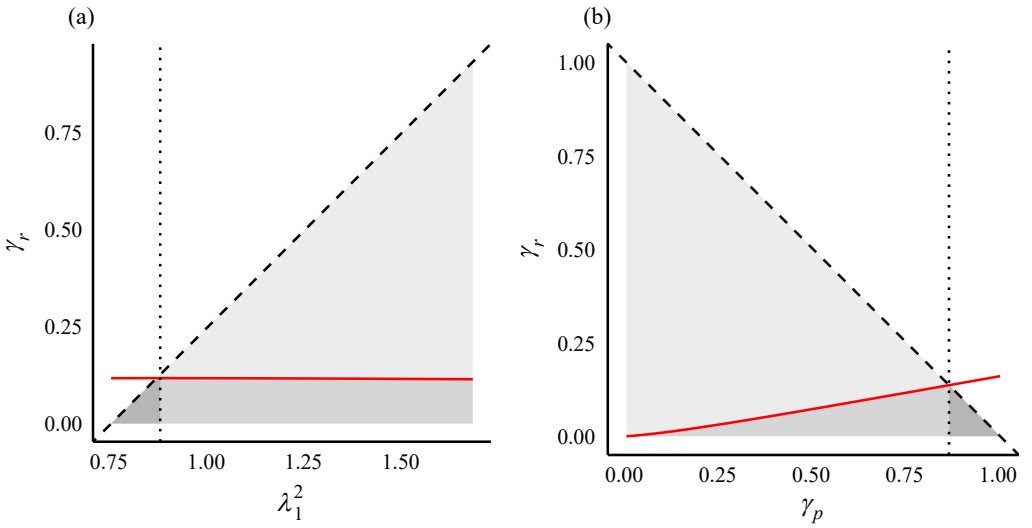

Fig. 2. Graphical illustration of the results of Theorem 3.

be asymptotically distinguishable from the noise, and yet the augmentation estimator is inconsistent. Similarly, statement (iii) states that, regardless of how strong the signal-to-noise ratio is, there always exists a large enough data-collection rate $\gamma_p > 0$ such that, although the signal remains asymptotically distinguishable from the noise, the augmentation yields an inconsistent estimate. Collectively, these results offer the following intuition: detecting weaker signals or operating in higher-dimensional regimes requires γ_r to be large enough to reveal a gap between the signal and noise. This aligns with the conjecture that the number of augmentations should be taken as large as possible, given (for finite p) in Radojičić et al. (2025). However, in some scenarios, such as when the signal is very weak or the dimensionality γ_p is very high, the theoretically required γ_r may fall outside the interval in which the signal is asymptotically identifiable (see Assumption 1), rendering consistent estimation impossible through original predictor augmentation; a phenomenon also observed in the numerical experiments of Radojičić et al. (2025). For illustration, see Fig. 2(a), which illustrates inconsistency regions for $\gamma_p = 0.75$, $\sigma^2 = 1$ and $d = 1$. The light grey area corresponds to the feasible region $\gamma_r + \gamma_p < \lambda_1^2$; see Assumption 1. The solid curve determines, for every λ_1, the value of γ_r^0 from Theorem 3(i), for which $\phi(1) = \phi(2)$. Therefore, in the darker grey area obtained as the intersection of the feasible area and the area below the solid curve, the augmentation estimator is inconsistent. In particular, the darkest grey area represents situations with feasibly strong signal (in the sense of Assumption 1), yet not strong enough so that predictor augmentation is inconsistent for every choice of γ_r for which $\lambda_1^2 > \gamma_p + \gamma_r$. Similarly, Fig. 2(b) illustrates inconsistency regions for $\lambda_1 = 1$, $\sigma^2 = 1$ and $d = 1$. Finally, we note that Theorem 3 covers only one possible case of inconsistent estimation, $\phi(d) > \phi(d+1)$, implying that inconsistency can also occur in further subregions of Fig. 2.

5. Consistent estimator of d in high-dimensional data

We next define an alternative to the augmentation function ϕ_n that retains its consistency for high-dimensional data. Our proposal has three key elements. (i) Unlike Luo & Li (2021), we estimate the signal eigenvalues directly from the original data (instead of from the augmented data). This leads to less biased estimates of λ_j. (ii) We further correct for the remaining bias in the estimated eigenvalues by applying an appropriately chosen transformation to them. (iii) We combine the eigenvalue and eigenvector information, not by summing them, but by adjusting the norms $\|u_{j,n,C}\|^2$ such that the jump from the signal to the noise becomes apparent in their plot.

As per item (i) above, we assume throughout this section that the eigenvalues $\tau_{j,n}$ have been estimated from the sample covariance matrix of the original data. In this case, the equivalent of

Theorem 2 holds with $\gamma_r = 0$. Define next the debiasing function $f_n : \mathbb{R} \to \mathbb{R}$ as

$$f_n(\tau) = \tfrac{1}{2}\{\tau - \sigma^2(1 + p_n/n)\} + \tfrac{1}{2}[\{\tau - \sigma^2(1 + p_n/n)\}^2 - 4\sigma^4(p_n/n)]_+^{1/2},$$

where $[a]_+ := \max\{0, a\}$ is the soft-thresholding function. By Theorem 2 and direct computation, we see that $f_n(\tau_{j,n}) \xrightarrow{\text{p}} \lambda_j$ for all $j = 1, \ldots, d$, showing that f_n allows for the unbiased estimation of the spikes. Using the debiased estimates $f_n(\tau_{j_n,n})$ and the eigenvector norms, we then construct

$$h_{j,n} := f_n(\tau_{j,n})\{f_n(\tau_{j,n}) + (p_n/n + r_n/n)\sigma^2\}\{f_n(\tau_{j,n}) + \sigma^2\}^{-1}\|u_{j,n,C}\|^2. \tag{2}$$

The reason for defining $h_{j,n}$ as in (2) is that the probability limit of $h_{j,n}$ is constant both for $j \in \{1, \ldots, d\}$ and for $j \in \{d + 1, \ldots, K\}$, where $K > d$ is a fixed constant. Theorem 4 below formalizes this and quantifies the size of the jump, which is always negative, between these two regions. Consequently, we can estimate d by locating this jump as $\arg\min_{j=1,\ldots,K}\{h_{j+1,n} - h_{j,n}\}$; see Corollary 1 below that states that this estimator is, unlike original PA, consistent under all possible high-dimensional regimes $\gamma_p, \gamma_r > 0$, under Assumption 1.

THEOREM 4. *Fix $K \in \mathbb{N}$ such that $K > d$. Then, under Assumption 1, we have $h_{j+1,n} - h_{j,n} \xrightarrow{p} 0$ for all $j = 1, \ldots, K$ with $j \neq d$. Also,*

$$h_{d+1,n} - h_{d,n} \xrightarrow{\text{p}} -\frac{\sigma^2\gamma_r^2}{(\sqrt{\gamma_p} + 1)(\gamma_p + \gamma_r)} < 0.$$

COROLLARY 1. *Fix any $K \in \mathbb{N}$ such that $K > d$. Then, under Assumption 1, $d_n := \arg\min_{j=1,\ldots,K}\{h_{j+1,n} - h_{j,n}\}$ satisfies $d_n \xrightarrow{p} d$.*

In Theorem 4 we search the true signal dimension within the finite set $\{1, \ldots, K\}$. This is practical since one, in any case, typically assumes that the signal of the data is captured by a small amount of factors. However, the search interval could also be widened and allowed to grow with n, in the same sense as discussed in Remark 1. Moreover, the explicit expression for the limit of $h_{d+1,n} - h_{d,n}$ in Theorem 4 could be used to tune parameter γ_r, so that this jump is maximized. If $\gamma_p > 0$, substituting $\gamma_r = c\gamma_p$ into the limit of $h_{d+1,n} - h_{d,n}$, we get

$$h_{d+1,n} - h_{d,n} \xrightarrow{\text{p}} T(\gamma_p, c, \sigma) := \frac{-\sigma^2c^2\gamma_p}{(\sqrt{\gamma_p} + 1)(1 + c)} < 0.$$

As $c, \gamma_p > 0$, it is easily shown that, for every $\gamma_p, \sigma^2 > 0$, $c \mapsto T(\gamma_p, c, \sigma)$ is strictly decreasing in c, implying that larger values of γ_r lead to a larger jump in the differences. The same is true if $\gamma_p = 0$, as then $|h_{d+1,n} - h_{d,n}| \xrightarrow{\text{p}} \sigma^2\gamma_r$. However, at the same time γ_r should be kept sufficiently far from the threshold implied by Assumption 1, since even close proximity to it might compromise the finite-sample performance of the method.

As in practice the noise variance σ^2 is unknown, we give a number of possible consistent estimators of it. The next result follows directly from Theorem 2 and we thus omit its proof.

COROLLARY 2. *Let $\tau_{j_n,n}$ be the j_nth eigenvalue of the sample covariance of $x_{1,n}, \ldots, x_{n,n}$. Then, for $j_n \in \{d + 1, \ldots, p_n\}$ such that $j_n/p_n \to 1 - q \in [0, 1]$, as $n \to \infty$ and under Assumption 1,*

$$\begin{aligned}\hat{\sigma}^2_{j_n} &:= g(\tau_{j_n,n}) \\ &= \begin{cases} \dfrac{\tau_{j_n,n}}{F^{-1}_{\gamma_p}(q)}, & \gamma_p \leqslant 1, \\ \dfrac{\tau_{j_n,n}}{\gamma_p F^{-1}_{\gamma_p^{-1}}(q)}, & \gamma_p > 1 \end{cases} \\ &\xrightarrow{\text{p}} \sigma^2,\end{aligned}$$

where F^{-1}_{γ} is as in Theorem 2.

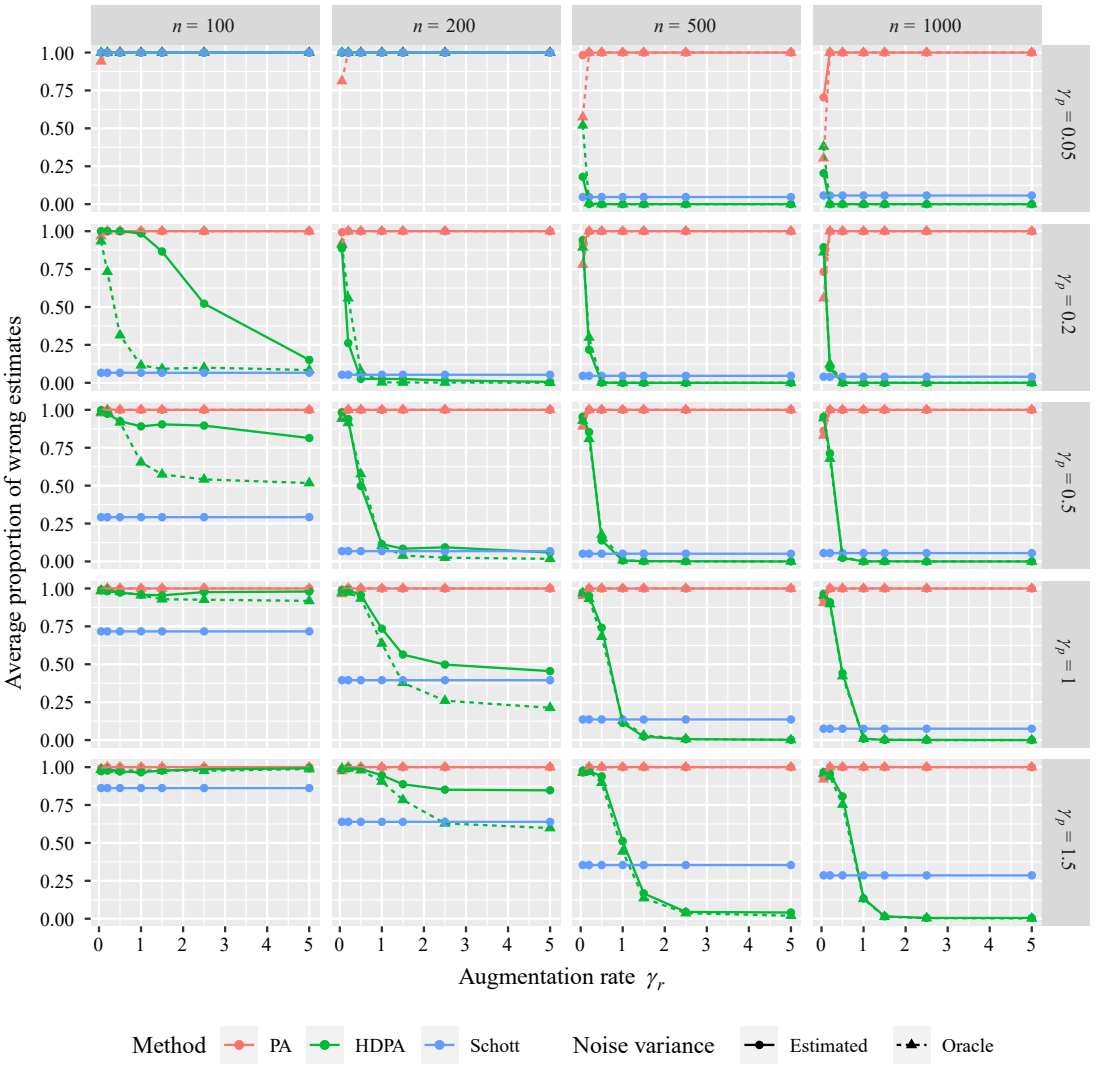

Fig. 3. Mean proportions of wrong estimates for $n \in \{100, 200, 500, 1000\}$, $p_n \in \{0.05n, 0.2n, 0.5n, n, 1.5n\}$.

6. Simulation

This simulation assesses the finite-sample accuracy of the proposed high-dimensional predictor augmentation (HDPA) and compares it against two competitors: the original PA of Luo & Li (2021) and the high-dimensional subsphericity-based estimator of Schott (2006). The data follow a Gaussian distribution $X \sim \mathcal{N}_{p_n}(0_{p_n}, \Sigma_{p_n})$, with $\Sigma_{p_n} = \mathrm{diag}(5, 4.8, \ldots, 3.2, 3, 0, \ldots, 0) + I_{p_n}$, giving latent dimension $d = 11$ and noise variance $\sigma^2 = 1$. We replicate $m = 1000$ datasets of size $n = 100, 200, 500, 1000$ and dimensionality $p_n = \gamma_p n$ for $\gamma_p = 0.05, 0.2, 0.5, 1, 1.5$. To study the effect of the augmentation dimension r_n, we estimate the latent dimension in all settings using $r_n = \gamma_r n$ for $\gamma_r = 0.05, 0.2, 0.5, 1, 1.5, 2.5, 5$.

To assess the effect of the noise variance estimation for PA and HDPA, we use the oracle $\sigma^2 = 1$, as well as the corrected median eigenvalue estimator $\hat{\sigma}^2_{\lfloor n/2 \rfloor}$ defined in Corollary 2. The mean proportions of wrong estimates for each method are given in Fig. 3, showing that HDPA outperformed PA as the latter fails to accurately estimate the dimension, and its performance does not improve with n. These findings are in line with the introductory discussion and Theorem 3. Furthermore, for $n \geqslant 500$,

HDPA estimates the dimension with very high accuracy. As HDPA relies on the limiting behaviour of the eigenvalues and eigenvectors of the sample covariances, it is no surprise that a larger sample size is needed to validate its consistency. Schott's method, based on successive hypothesis testing, can at best have a 0.05 proportion of wrong estimates due to the 0.05 significance level. While letting the level depend on n could improve this, it would be impractical and lack finite-sample guarantees. In accordance with Theorem 4, HDPA gets more accurate with increasing γ_r. For large enough n, we observe no difference in the performance of HDPA regardless of whether we use the oracle noise variance or its estimate. To test the sensitivity of the HDPA to the violation of the Gaussianity assumption, we repeated the simulation in five additional settings. These, and a real data example, are given in the Supplementary Material.

ACKNOWLEDGEMENT

The work of Virta was supported by the Research Council of Finland (335077, 347501, 353769). The work of Radojičić was funded by the Austrian Science Fund (FWF) [10.55776/I5799].

SUPPLEMENTARY MATERIAL

The Supplementary Material contains proofs of the main results and additional simulations.

REFERENCES

ALAOUI, A. E., KRZAKALA, F. & JORDAN, M. (2020). Fundamental limits of detection in the spiked Wigner model. *Ann. Statist.* **48**, 863–85.

BERNARD, G. & VERDEBOUT, T. (2024). Power enhancement for dimension detection of Gaussian signals. *Statist. Sinica* **34**, 2161–82.

BLOEMENDAL, A., KNOWLES, A., YAU, H.-T. & YIN, J. (2016). On the principal components of sample covariance matrices. *Prob. Theory Rel. Fields* **164**, 459–552.

FAN, J., FAN, Y. & LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Economet*. **147**, 186–97.

JAGANNATH, A., LOPATTO, P. & MIOLANE, L. (2020). Statistical thresholds for tensor PCA. *Ann. Appl. Prob*. **30**, 1910–33.

JOHNSTONE, I. M. & LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc*. **104**, 682–93.

LUO, W. & LI, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–87.

LUO, W. & LI, B. (2021). On order determination by predictor augmentation. *Biometrika* **108**, 557–74.

MUKHERJEE, S. S. (2025). Consistent model selection in the spiked Wigner model via AIC-type criteria. *arXiv:* 2307.12982v2.

NORDHAUSEN, K., OJA, H. & TYLER, D. E. (2022). Asymptotic and bootstrap tests for subspace dimension. *J. Mult. Anal*. **188**, 104830.

RADOJIČIĆ, U., LIETZÉN, N., NORDHAUSEN, K. & VIRTA, J. (2025). Order determination for tensor-valued observations using data augmentation. *J. Comp. Graph. Statist*., doi: 10.1080/10618600.2025.2500977.

SCHOTT, J. R. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *J. Mult. Anal*. **97**, 827–43.

YAO, J., ZHENG, S. & BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge: Cambridge University Press.

[*Received on* 9 *February* 2025. *Accepted on* 3 *July* 2025]