Annals of Science and Inquiry
Vol. 2 No. 11 (2026)

A High-Variance Component Can Erase Monitoring Contrasts

Graham Mills · Wesley King · Timothy W. Clark

Published: November 26, 2026 | Article 1135

Abstract

The leading component explained 95.0% of training variance across twelve Beijing monitoring sites. Yet its covariance approximation assigned less than half the observed holdout variance to each of the 66 pairwise station differences. The apparent contradiction comes from the quantity being measured. The component retained broad fluctuations shared across the network, and much of that movement canceled in a site difference. With four components, restoring the omitted variance separately for each site reduced the median absolute log error of the contrast variance from 0.440 to 0.149 on later days. The same correction increased the error in summer, when the earlier full covariance already overstated most contrast variances. Comparing compressed estimates with the earlier uncompressed estimate helped distinguish loss of spatial detail from changes over time.

A High-Variance Component Can Erase Monitoring Contrasts

Abstract

The leading component explained 95.0% of training variance across twelve Beijing monitoring sites. Yet its covariance approximation assigned less than half the observed holdout variance to each of the 66 pairwise station differences. The apparent contradiction comes from the quantity being measured. The component retained broad fluctuations shared across the network, and much of that movement canceled in a site difference. With four components, restoring the omitted variance separately for each site reduced the median absolute log error of the contrast variance from 0.440 to 0.149 on later days. The same correction increased the error in summer, when the earlier full covariance already overstated most contrast variances. Comparing compressed estimates with the earlier uncompressed estimate helped distinguish loss of spatial detail from changes over time.

Which variation should survive compression?

A network covariance matrix can describe both common variation and differences across otherwise similar locations. The variance of a daily average summarizes how the monitored sites move together. The variance of a difference between two sites measures their local divergence. A spatial component that has a similar loading at every site is useful for the first quantity and contributes little to the second. If the goal is to reduce the size of a representation used in an uncertainty calculation, a high percentage of total retained variance is therefore an incomplete criterion. It describes the covariance on its eigenvectors, while the downstream directions may be fixed by an application.

Principal-component analysis has long provided an ordering of directions by explained variation (see Pearson, 1901 and Hotelling, 1933). Modern expositions describe its geometry and applications (Jolliffe and Cadima, 2016) (Jolliffe, 2002) (Greenacre et al., 2022). The rank criterion remains appropriate when squared reconstruction error over the full variable vector is the target. For an application using particular linear combinations, accuracy instead depends on how the estimated matrix acts on those directions. Jiang et al. (2026) address the role of dimensionality reduction in moment-based distributionally robust optimization. We examine an empirical input question suggested by that work. When a network covariance is compressed, does adding back residual marginal variance make spatial contrast variances estimated in one period more accurate in a later period?

The answer is relevant to a user who has a compact covariance representation and needs to compare two sites using the observations already available across the network. It is also an opportunity to distinguish distortion introduced by rank truncation from a change in the measured covariance itself. We therefore keep a full earlier covariance in the comparison. A factor model can look worse than the full matrix because it drops differences among sites. It can also look better through partial shrinkage when the earlier full estimate overstates later variation. Neither possibility follows directly from the share of retained training variance.

Data and chronological holdout

The public Beijing Multi-Site Air Quality release contains hourly environmental observations from twelve locations. We use the recorded PM2.5 concentration from March 2013 through February 2017, one variable in physical units at each station. The source is the UCI data deposit (Chen, 2017). Zhang et al. (2017) describe the broader monitoring investigation. Our twelve-site archive has a narrower purpose. We analyze covariance estimation for the observed network and do not attribute changes in concentrations or variance to particular events or interventions.

The source contains 420,768 station-hours and 8,739 missing values in the concentration field. We group stations by the date and hour given in the release. Each daily vector uses only hours for which all twelve stations have valid values, so every pairwise difference uses the same hours on that day. The main sample requires at least 18 common hours, retaining 553 training days and 563 holdout days. The earlier period runs from March 1, 2013 through February 28, 2015. The later period begins March 1, 2015 and ends February 28, 2017. Each half contains two complete seasonal cycles before filtering. A stricter screen requires all 24 common hours and retains 303 training days and 220 holdout days.

Using the earlier dates only for fitting makes the comparison an out-of-period exercise. Work on forecasting evaluation distinguishes this kind of holdout from arbitrary record splits (Tashman, 2000) (Cerqueira et al., 2020) (Bergmeir et al., 2018). We repeat the analysis after selecting each meteorological season in both halves. The retained dates in a given season form their own samples and the covariance is fitted anew within that season. This is a fixed sensitivity analysis rather than a tuning rule. In particular, no observed holdout error is used to choose the rank of the model to be evaluated.

Representations of the omitted variance

Let the earlier sample covariance be the sum of contributions ordered by its eigenvalues. We compare ranks one, two and four. The truncated representation retains the first selected contributions. It gives the best low-rank reconstruction in the usual spectral geometry, and the variances along discarded eigenvectors are zero under this representation. The common coordinate scale means that no site receives different weight from rescaling its physical units. Choices about centering and scaling would alter the spectral comparison (Bro and Smilde, 2003). Independently centered earlier and later samples let the study compare dependence around each period's own mean rather than deviations from a fixed citywide level.

Two residual corrections use only the training matrix. In the isotropic correction, each site receives an equal diagonal addition equal to the mean of the discarded marginal variances. In the site-specific correction, each receives its own discarded marginal variance. Both additions are positive semidefinite. The site-specific correction reproduces all training marginal variances exactly and leaves discarded off-diagonal covariances absent. Neither is assumed to reconstruct the full omitted covariance. The isotropic idea is related to residual assumptions in probabilistic PCA (Tipping and Bishop, 1999), and low-rank models combined with an estimated residual matrix appear in the covariance literature (Fan et al., 2008) (Fan et al., 2013). The present calculations are simpler. We add known functions of the training residual and do not fit a likelihood or infer a latent generative process.

All 66 unordered pair differences are assigned equal weight. For each, we evaluate the variance implied by the earlier representation and divide it by the empirical variance in later data. A ratio below one means underestimation. The main summary is the median absolute logarithm of these ratios, so estimating twice as much and estimating half as much have the same error. We also count ratios within plus or minus 25% of one and report the estimate for the equal-weight network mean as a separate direction. These counts and medians describe a finite set of directions sharing common sites. Treating 66 overlaps as independent data would overstate the precision of the comparison.

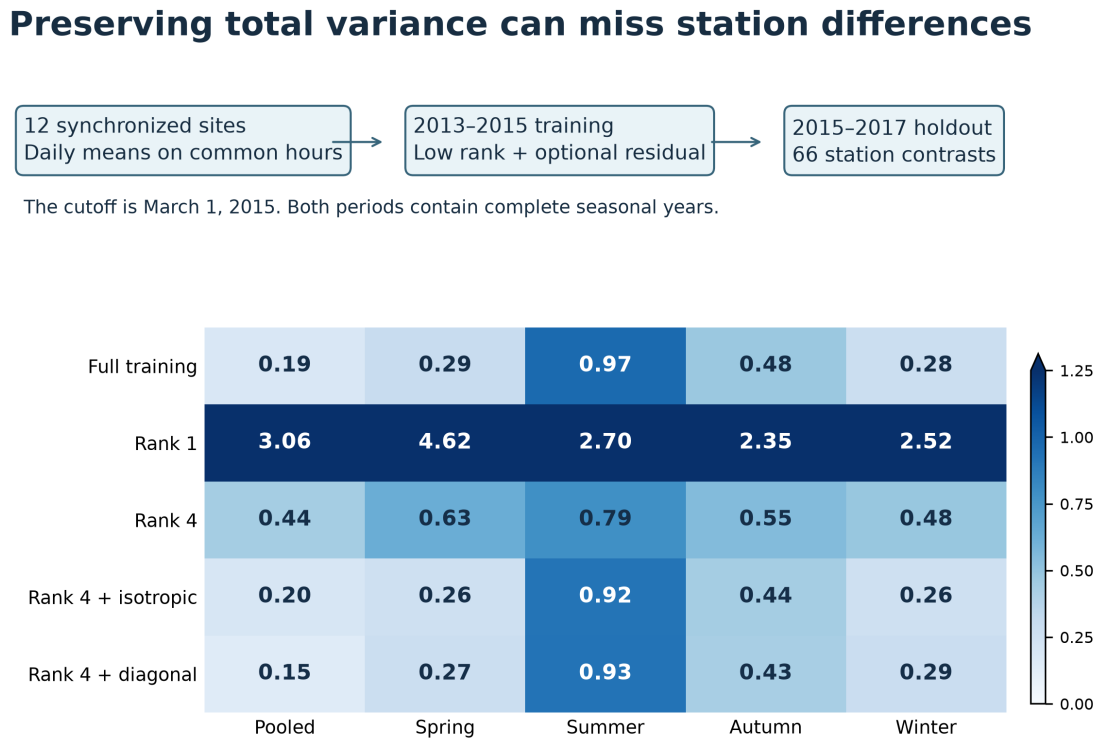

Figure 1. Each covariance representation uses an earlier period or its season-specific subset. The displayed values evaluate 66 station differences against observed variances in the corresponding later sample.

Total variance and spatial differences tell different stories

In the pooled training sample, the leading component retains 95.02% of total covariance trace. Its estimate for the held-out network mean variance is 0.767 times the later observed variance, almost exactly the ratio produced by the full earlier covariance. For pairwise contrasts the situation differs. The median estimate from the same component is 0.047 times the later variance and all 66 estimates fall below half the later value. The spectral representation retains the strongest common pattern while assigning very little variance to many differences. This is a case in which an apparently adequate representation of a regional statistic has consistently inadequate scale for comparisons between sites.

Table 1. Pooled chronological comparison on days with at least 18 common hours. Trace fractions describe the low-rank part before either residual is restored.

Earlier covariance model	Explained trace before correction	Median pairwise ratio	Median absolute log error	Ratios within 25%, out of 66	Network mean variance ratio
Full sample covariance	100.00%	1.008	0.186	44	0.767
Rank 1	95.02%	0.047	3.058	0	0.767
Rank 4	98.63%	0.670	0.440	17	0.767
Rank 4 plus isotropic residual	98.63%	0.982	0.203	41	0.768
Rank 4 plus site-specific residual	98.63%	0.970	0.149	42	0.768

Rank-four correction improves the pooled pairwise error and has little effect on the variance estimate for the network mean. Compared with the full sample covariance, the site-specific estimate has a smaller median absolute log error, while the full sample covariance places two more pairs within the stated tolerance. Those two summaries respond differently to the distribution of errors. The comparison shows an improvement in the median error over directions, but does not establish uniform dominance. Covariance shrinkage methods explicitly depend on the structure they target and the loss used to measure accuracy (Ledoit and Wolf, 2004) (Daniels and Kass, 2001) (Bickel and Levina, 2008). Restoring a diagonal is a narrow structural choice. In the pooled comparison, its usefulness depends on testing the intended linear contrasts rather than selecting it only by total trace.

The intermediate rank confirms that residual correction and component count should be considered together. Rank two retains 97.40% of the earlier trace and produces a pooled contrast error of 0.995. Its isotropic residual reduces that to 0.203, while its site-specific residual reduces it to 0.233. At rank four the order between the two residuals reverses. The pooled evidence therefore supports estimating some omitted variance when comparing locations. It does not make a site-specific residual the more accurate choice at every rank. Even though that representation preserves all twelve earlier site variances, the variance of a difference also contains twice the between-site covariance. Neither diagonal addition provides that missing covariance.

What the seasons reveal

Summer shows why a low-rank correction can address a different problem from variation across time. The median primary pairwise ratio for the full earlier summer covariance is 2.638. Even with a site-specific diagonal correction, the rank-four representation has a median ratio of 2.526. The later summer covariance is smaller along many of the selected contrasts. Simply restoring variance discarded by PCA cannot align a large earlier variance to a smaller later variance. The uncorrected rank-two summer approximation has smaller median absolute log error than either corrected rank-four estimate, but its lower magnitude is partly the result of throwing variance away. The pooled comparison provides no general rule that the same amount of truncation will help in summer.

The winter result differs. The full covariance there has a median variance ratio of 0.787, and the site-specific rank-four estimate has a ratio of 0.771. Both miss some later variation. In spring the median ratio of the corrected representation is 1.145, against 0.637 for uncorrected rank four. Autumn shows a similar movement toward one, from 0.732 to 1.146. Both corrections use the same mathematical rule in every season. Their effect in later data changes with the distribution of spatial contrasts. This is consistent with the statistical distinction between estimating a contemporaneous covariance and modeling a varying factor process (Forni and Lippi, 2011). No dynamic factor parameters are fitted here.

Requiring all 24 common hours changes which days contribute, particularly in the holdout. The pooled rank-one median log error is still 3.031. Rank-four error moves from 0.514 without residuals to 0.237 with site-specific residuals, close to the full earlier covariance error of 0.228. This sensitivity supports the contrast-loss finding under a stronger data-coverage rule. It does not establish performance on days with too few overlapping hours. Multivariate monitoring literature also distinguishes representation choices from the decision rules that later use them (Qin, 2003) (Ku et al., 1995). We do not set alarm thresholds or interpret the error counts as rates of operational detection.

The common-hour rule also limits the effect of uneven within-day availability. Subtracting station means formed over different hours would allow a shared fluctuation during a missing hour to appear as a spatial contrast. Keeping the same daily hours for each station avoids that particular mismatch. The resulting daily averages remain conditional on what the shared masks include. A day with 18 observed hours may

lack a persistent part of the diurnal cycle, whereas a fully observed day represents the reported 24-hour sequence. Comparing the two selection rules helps with the dimension-reduction question because all candidate methods use the same observations within each rule. It does not recover any unreported hourly values.

Limits of a finite moment representation

The covariance and paired ratios represent a conditional archive, not spatial exposure at an arbitrary location. Station differences were used because they have stakeholder-independent definitions and are easily reproduced from the archived series. A different task could select contrasts by distance, socioeconomic boundary, or another known direction. A rank chosen for one set need not perform similarly for another. Work on factor-number selection supplies ways to estimate dimensionality under explicit modeling assumptions (Bai and Ng, 2002), and robust PCA illustrates what could change if occasional large observations drive the fitted directions (Hubert et al., 2005). Neither is an adjustment applied to the current data.

Moment-based robust optimization uses a mean and covariance along with assumptions that govern the uncertainty set (Delage and Ye, 2010) (Wiesemann et al., 2014). Chance-constrained and moment-inequality formulations show how joint moments influence bounds for particular directions (Zymler et al., 2013) (Bertsimas and Popescu, 2005). High total explained variance may leave the covariance too small in a direction the decision uses. Keeping the training diagonals addresses this particular source of shrinkage without protecting against errors in cross-covariance or temporal transfer. The two reference scales, full covariance and rank truncation without residuals, help an analyst locate which failure could be addressed by that simple change.

The frozen source records are licensed CC BY 4.0. The companion script records the archive hash, station names, period boundary, common-hour types and all paired variance ratios. The figures and counts can be replayed from the same dated observations. Sample covariance uses the standard denominator for a centered empirical moment. Discussions of multivariate sampling distributions explain why more formal inference requires assumptions beyond such a computation (Hotelling, 1931), and using fixed central summaries avoids discrepancies from competing quantile definitions (Hyndman and Fan, 1996). Missing-data work emphasizes that selection can change an observed sampling frame (Rubin, 1976). The data permit a practical, more limited conclusion. When the intended statistic compares sites, its variance is more informative about the adequacy of compression than the single network-wide fraction of trace retained.

References

Bai, J., & Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, *70*(1), 191-221. https://doi.org/10.1111/1468-0262.00273

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70-83. https://doi.org/10.1016/j.csda.2017.11.003

Bertsimas, D., & Popescu, I. (2005). Optimal Inequalities in Probability Theory: A Convex Optimization Approach. *SIAM Journal on Optimization*, *15*(3), 780-804. https://doi.org/10.1137/s1052623401399903

Bickel, P. J., & Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, *36*(6), 2577-2604. https://doi.org/10.1214/08-aos600

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, *17*(1), 16-33. https://doi.org/10.1002/cem.773

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, *109*(11), 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

Chen, S. (2017). *Beijing Multi-Site Air Quality* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5RK5G

Daniels, M. J., & Kass, R. E. (2001). Shrinkage Estimators for Covariance Matrices. *Biometrics*, *57*(4), 1173-1184. https://doi.org/10.1111/j.0006-341x.2001.01173.x

Delage, E., & Ye, Y. (2010). Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, *58*(3), 595-612. https://doi.org/10.1287/opre.1090.0741

Fan, J., Fan, Y., & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, *147*(1), 186-197. https://doi.org/10.1016/j.jeconom.2008.09.017

Fan, J., Liao, Y., & Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(4), 603-680. https://doi.org/10.1111/rssb.12016

Forni, M., & Lippi, M. (2011). The general dynamic factor model: One-sided representation results. *Journal of Econometrics*, *163*(1), 23-28. https://doi.org/10.1016/j.jeconom.2010.11.003

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, *2*(3), 360-378. https://doi.org/10.1214/aoms/1177732979

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417-441. https://doi.org/10.1037/h0071325

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, *47*(1), 64-79. https://doi.org/10.1198/004017004000000563

Hyndman, R. J., & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, *50*(4), 361-365. https://doi.org/10.1080/00031305.1996.10473566

Jolliffe, I. T. (2002). Principal Component Analysis. *Springer Series in Statistics*. https://doi.org/10.1007/b98835

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Ku, W., Storer, R. H., & Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 179-196. https://doi.org/10.1016/0169-7439(95)00076-3

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365-411. https://doi.org/10.1016/s0047-259x(03)00096-4

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572. https://doi.org/10.1080/14786440109462720

Jiang, S., Cheng, J., Pan, K., & Shen, Z. M. (2026). Optimized Dimensionality Reduction for Moment-Based Distributionally Robust Optimization. *Operations Research*, *74*(3), 1648-1669. https://doi.org/10.1287/opre.2023.0645

Qin, S. J. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, *17*(8-9), 480-502. https://doi.org/10.1002/cem.800

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437-450. https://doi.org/10.1016/s0169-2070(00)00065-0

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *61*(3), 611-622. https://doi.org/10.1111/1467-9868.00196

Wiesemann, W., Kuhn, D., & Sim, M. (2014). Distributionally Robust Convex Optimization. *Operations Research*, *62*(6), 1358-1376. https://doi.org/10.1287/opre.2014.1314

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., & Chen, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *473*(2205), 20170457. https://doi.org/10.1098/rspa.2017.0457

Zymler, S., Kuhn, D., & Rustem, B. (2013). Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, *137*(1-2), 167-198. https://doi.org/10.1007/s10107-011-0494-7