Journal of Data Efficiency
Volume 1 · Number 1 · 2026

Ten Times the Fitting Images, Slightly Worse Completion: A Fixed-Rank MNIST Study

Bradley Nelson[1] · Gabriel Stone[2] · James Harrison[3]

[1] Sweden · [2] Finland · [3] Denmark

Original publication: October 24, 2026 | Article 608

Abstract

Would collecting more complete examples necessarily help a small linear model fill a missing image region? On 10,000 official MNIST test images, we compared rank-32 reconstruction bases fitted to nested samples of 500 and 5,000 training images. The larger fit increased mean squared error inside a common eight-by-eight central mask from 0.138980 to 0.142867. The paired increase was 0.003886, with a conditional image-level 95% interval of 0.002472 to 0.005301. Both fits improved on their respective mean-image baselines. Digit-level averages also moved in different directions. This one-draw result concerns a particular basis, inference penalty, and mask; it does not show that additional data generally harm restoration.

Public handwritten digits with an artificial central mask; a comparison of two linear reconstruction fits, rather than a test of a neural restoration system.

Ten Times the Fitting Images, Slightly Worse Completion: A Fixed-Rank MNIST Study

Public handwritten digits with an artificial central mask; a comparison of two linear reconstruction fits, rather than a test of a neural restoration system.

Abstract

Would collecting more complete examples necessarily help a small linear model fill a missing image region? On 10,000 official MNIST test images, we compared rank-32 reconstruction bases fitted to nested samples of 500 and 5,000 training images. The larger fit increased mean squared error inside a common eight-by-eight central mask from 0.138980 to 0.142867. The paired increase was 0.003886, with a conditional image-level 95% interval of 0.002472 to 0.005301. Both fits improved on their respective mean-image baselines. Digit-level averages also moved in different directions. This one-draw result concerns a particular basis, inference penalty, and mask; it does not show that additional data generally harm restoration.

1. A data requirement is a claim about an entire procedure

The question of how much training data a restoration method needs is meaningful only after its representation, fitting rule, and evaluation task have been specified. Additional observations may estimate a population covariance more accurately, but an improved covariance estimate need not minimize error for every missing-region prediction problem. A representation chosen to explain variation across complete images is not chosen directly to reconstruct the center from the surrounding pixels. The distinction supplies a concrete reason to examine, rather than assume, the benefit of a larger fitting sample.

We ask whether a tenfold increase in the number of complete training images reduces central-block reconstruction error when the retained rank and test-time inference rule remain fixed. The directional proposition is that the larger sample should have lower average masked error on the same official test images. Its alternative is simply a zero or positive average difference. No claim about all possible training-set sizes or restoration architectures follows from either outcome. The comparison uses one deliberately limited linear procedure so that the meaning of the data budget remains visible.

Work on character restoration, including Meng et al., 2025, gives practical context for thinking about examples available to an image-completion system. Our observations are contemporary handwritten digits, and our executed method is a covariance basis with regularized latent inference. We do not evaluate oracle-bone rubbings, reproduce that paper's architecture, or measure whether a reader can identify a restored character. The connection is the question of evidence needed for a data-requirement statement, not a transfer of model performance between scripts.

2. Why a linear representation can answer a useful narrow question

The principal-axis perspective in Pearson, 1901 and the latent-variable treatment in Tipping and Bishop, 1999 provide foundations for distinguishing a complete-data representation from a prediction task. We use ordinary covariance eigenvectors, not probabilistic principal-component maximum likelihood or an expectation-maximization fit. The leading directions describe complete-image variation in the selected training sample. Their usefulness for inferring an unobserved center depends on which combinations of those directions remain identifiable from the visible image boundary.

The penalty applied to latent coefficients is motivated by the stabilization idea associated with ridge regression in Hoerl and Kennard, 1970. Here the fitted image basis has thirty-two columns, while the observation mask determines which rows enter the latent least-squares problem. Keeping the penalty numerically equal across the two data budgets makes that inference rule reproducible. It does not establish that this penalty is optimal for either basis. A representation with less convenient conditioning on the visible pixels might require a different penalty even when its full-image variance summary looks favorable.

There are many more expressive routes to restoration. Denoising representation learning in Vincent et al., 2008, context prediction in Pathak et al., 2016, and mask-aware convolutional approaches in Liu et al., 2018 and Yu et al., 2018b frame different ways to learn missing-content structure. None supplies generated images or fitted weights to this study. The present comparison is useful precisely because it can be inspected without interpreting an additional neural training trajectory. It cannot be used to rank those approaches or estimate their sample efficiency.

3. The official image partitions and the nested fitting samples

We read the four compressed MNIST image and label files, verify their IDX headers and payload lengths, and retain the official separation of 60,000 training images and 10,000 test images. Each image is twenty-eight pixels square. Unsigned integer intensities are divided by 255, producing the zero-to-one scale used for both fitting and loss calculation. The document-recognition setting in Lecun et al., 1998 identifies the source task; our endpoint is reconstruction error rather than its usual digit classification outcome.

A random permutation of the training indices is generated with seed 20261107. The smaller fit uses its first 500 indices and the larger fit its first 5,000. Consequently every image in the smaller sample also occurs in the larger one. This nested construction avoids comparing completely unrelated samples at the two budgets. It still represents only one permutation. A different initial set of 500 images could change the contrast, and the confidence interval reported below contains no repeated-training-sample component.

Training and test labels are unnecessary for estimating either basis or any test-image latent coefficient. The test labels are retained only for descriptive averages after the reconstruction calculation. We do not choose the fitting samples to balance digits, select a favorable digit subset, or tune the procedure on the test outcome. The comparison changes both the sample mean and covariance basis because both are estimated independently at each budget. Describing it as a pure intervention on eigenvectors alone would therefore be inaccurate.

4. Complete-image fitting, visible-pixel inference

For each fitting sample we estimate its pixelwise mean and centered covariance matrix. The thirty-two eigenvectors associated with the largest eigenvalues form the reconstruction basis. They are checked for orthonormality to a numerical tolerance of one part in ten billion. Retaining a fixed number of directions gives the two representations the same latent dimension. It does not force them to span the same image subspace, allocate variance similarly across digit classes, or predict the masked pixels with the same bias.

Every official test image receives exactly the same artificial mask: zero-based rows ten through seventeen and columns ten through seventeen. These sixty-four central pixels are withheld from latent inference. The remaining 720 observed intensities, centered using the appropriate training mean, determine a regularized least-squares estimate of the thirty-two latent coefficients. The penalty is 0.01. We verify the corresponding normal equations numerically, then reconstruct the central pixels from the estimated coefficients and that fit's central mean values.

Predictions are clipped to the admissible intensity interval before scoring. The outcome is the average squared prediction error over the sixty-four hidden pixels, calculated separately for each test image. Clipping, masking, rank, and penalty are identical across fits. An additional baseline fills the masked region using only each sample's mean image. This baseline helps distinguish the usefulness of visible-pixel conditioning from a change in the mean. It does not introduce another trained architecture or use the missing test intensities to choose coefficients.

5. The model diagram follows the information boundary

Large-mask completion in Suvorov et al., 2021 and diffusion-based inpainting in Lugmayr et al., 2022 emphasize that the missing region and reconstruction mechanism are material parts of a task. The image-prior setting in Ulyanov et al., 2017 further illustrates why a restoration method's source of information must be specified. In this experiment complete training images provide the means and bases; only observed test pixels provide the test-specific latent coefficients. The hidden test pixels enter scoring after predictions have been finalized.

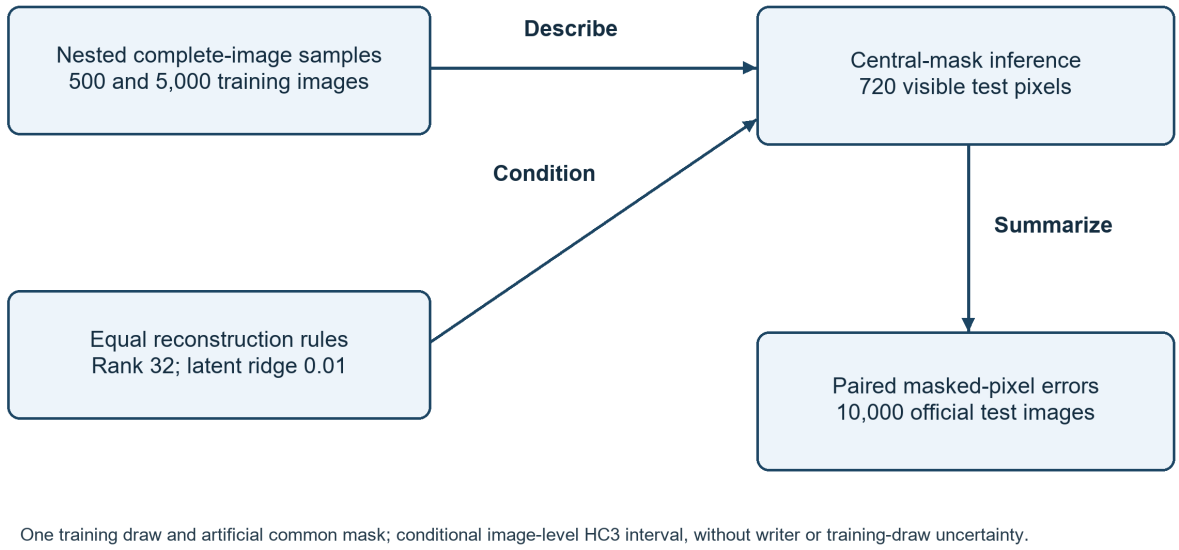


Figure 1. Two training-only covariance fits feed an identical visible-pixel inference rule; paired masked errors are compared on the official test partition.

State-space vision work in Guo et al., 2024 and Liu et al., 2024 is relevant to the broader choice of image representation, but no state-space model is fitted in our calculation. The figure therefore contains neither a neural encoder nor a supposed semantic-quality module. Its main purpose is to make the restriction on test information legible. Adding an architectural label to these boxes would imply an experiment that the data and computations do not contain.

6. Paired estimation and the limited meaning of its interval

For each test image we subtract the smaller-fit masked error from the larger-fit masked error. An intercept-only ordinary least-squares regression estimates the average of these ten thousand paired differences. Its coefficient is therefore an error contrast in squared normalized intensity units. A positive coefficient favors the smaller fit. There are no digit controls, image-quality adjustments, or selected covariates. Pairing already ensures that the two errors refer to the same underlying image and missing region.

We report HC3 standard errors, within the heteroskedasticity-consistent inference tradition of White, 1980, with a normal-reference 95% interval. The fitted intercept-only R-squared is effectively zero, as expected for a mean with no explanatory regressors. This fit statistic is not the explained variance of the image reconstruction model. The large sample size describes test images, while the number of realized training-sample draws remains one. These are different replication dimensions and cannot be substituted for each other.

Table 1. Conditional mean difference in masked-pixel squared error, larger fitting sample minus smaller fitting sample.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.003886 (0.000722)	[0.002472, 0.005301]	10000
M1: fit	R-squared = 0.0000	-	-	10000

Model labels: M1 = Five-thousand minus five-hundred image basis MSE.

Writer identities are unavailable in these IDX files. The dependence considerations discussed in Cameron and Miller, 2015 thus cannot be resolved by an invented writer cluster label. The interval treats image-level paired observations as its available sampling units and may be too narrow if substantial shared-writer dependence remains. Structural separation concerns in Roberts et al., 2017 also apply to future replications. Neither this covariance correction nor the untouched test partition quantifies how much the result would vary across newly sampled sets of training writers.

7. What changed when the fitting budget increased

Average masked error is 0.138980414 for the 500-image fit and 0.142866552 for the 5,000-image fit. Their difference is 0.003886138, approximately 2.8 percent of the smaller fit's average error. The HC3 standard error is 0.000721671, giving an interval from 0.002471689 to 0.005300586. Under this conditional calculation the increase is distinguishable from zero. It is a modest deterioration in a specified numerical endpoint, not evidence that collecting data is generally counterproductive.

The mean-only baselines yield errors of 0.184519116 and 0.183970892 at the smaller and larger budgets respectively. The larger mean is slightly better by that endpoint, while the larger covariance reconstruction is worse. Both covariance reconstructions nevertheless have lower average error than filling from their own means. These facts rule out the simplistic reading that visible pixels contributed nothing. They also show why a budget comparison should describe the fitted prediction procedure rather than infer its behavior from the mean image alone.

The retained directions account for 0.757190 of complete training variance in the smaller sample and 0.745122 in the larger sample. These are summaries computed on different fitting samples, so their difference is not a held-out model-quality contrast. A higher retained fraction in the smaller sample could reflect its more limited empirical variation. The numbers do not establish the mechanism of the masked-error increase. In particular, no mediation analysis identifies a variance fraction as the cause of worse prediction.

8. Digit averages resist a universal explanation

The larger basis improves the average error for digits zero, one, and five, with differences of about -0.000482, -0.005597, and -0.004671. It increases error for the other digit labels. The largest descriptive increases occur for seven, at 0.018606, and eight, at 0.013316. These subgroup averages show that the pooled contrast is not a common shift imposed on every class. They do not identify which strokes were recovered, which examples remained readable, or why a particular basis direction helped a particular digit.

No subgroup hypothesis tests are used to choose a preferred sample size. Selecting rank or penalty after inspecting these test differences would alter the evaluation, raising the model-selection concerns described in Varma and Simon, 2006. Likewise the influence-analysis perspective of Cook, 1977 would motivate inspecting highly influential individual errors, but no inconvenient observations are removed here. The digit summaries remain a descriptive decomposition of the reported pooled result, with their original test composition preserved.

Agreement concepts in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 distinguish reproducibility or interchangeability from a simple difference in means. We do not estimate an agreement statistic for the two fitted procedures. The paired mean answers an aggregate error question and leaves the distribution of individual prediction disagreements largely uncharacterized. It would be inappropriate to declare the two reconstructions interchangeable merely because their pooled losses differ by only a few thousandths.

9. Pixel prediction does not settle character meaning

The structural and perceptual motivations associated with Wang et al., 2004 and Zhang et al., 2018 explain why squared intensity loss should not be renamed semantic fidelity. Neither of those metrics is computed here, and neither a recognition model nor human readers score the reconstructed digits. A wrong central stroke can matter greatly for identity even when it occupies a small number of pixels. Conversely a numerically imperfect intensity value need not change what a reader sees. The present result is restricted to its directly calculated endpoint.

The artificial central mask also differs from naturally missing evidence. The missing-data framework in Little and Rubin, 2002 makes the mechanism a substantive part of an interpretation, while measurement-error reasoning in Carroll et al., 2006 directs attention to the observed representation. A scanner artifact, torn document, or uneven rubbing could have a different geometry and relationship to the visible image. We have not observed such damage or estimated a model of it. The experiment uses known complete references to isolate one reproducible completion task.

Interpretation guided by Wasserstein and Lazar, 2016 emphasizes the size and scope of the contrast rather than treating a small tail probability as a general law. The reproducibility concerns in Ioannidis, 2005 are especially relevant to one favorable or unfavorable training draw. A sensible next study would repeat the nested sampling, vary masks using a declared design, and select hyperparameters without using the final test outcomes. Those operations have not been executed for this report and must not be described as robustness checks already passed.

10. What this result contributes to a data-collection decision

For this fixed-rank procedure, ten times as many fitting images did not buy lower error on the selected central-block task. That finding directs attention to representation and conditioning, while leaving their causal contribution unresolved. If a restoration project faces a data-acquisition decision, a learning-curve study should use the actual missingness pattern and quality endpoint it intends to improve. A complete-image variance statistic alone cannot answer whether additional examples will improve a particular reconstruction operation.

Documentation principles in Gebru et al., 2021 motivate recording the official partitions, index permutation, normalization, mask coordinates, penalty, clipping rule, and absence of writer identifiers. Preservation under Wilkinson et al., 2016 includes the raw compressed sources, per-image paired errors, digit summaries, and the executable numerical procedure. These records make the reported reversal inspectable. They do not confer independent publication value or cultural-heritage validity; those judgments require additional scientific and domain review beyond this small public-data comparison.

References

1. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
2. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. https://doi.org/10.1109/5.726791.
3. Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
4. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
5. Tipping, Michael E.; Bishop, Christopher M.. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/1467-9868.00196.
6. Liu, Yue; Tian, Yunjie; Zhao, Yuzhong; Yu, Hongtian; Xie, Lingxi; Wang, Yaowei; Ye, Qixiang; Jiao, Jianbin; Liu, Yunfan. (2024). VMamba: Visual State Space Model. arXiv:2401.10166. https://arxiv.org/abs/2401.10166.
7. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
8. Guo, Hang; Li, Jinmin; Dai, Tao; Ouyang, Zhihao; Ren, Xudong; Xia, Shu-Tao. (2024). MambaIR: A Simple Baseline for Image Restoration with State-Space Model. arXiv:2402.15648. https://arxiv.org/abs/2402.15648.
9. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.
10. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
11. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.
12. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
13. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.

14. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

15. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.

16. Vincent, Pascal; Larochelle, Hugo; Bengio, Yoshua; Manzagol, Pierre-Antoine. (2008). Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning - ICML '08. https://doi.org/10.1145/1390156.1390294.

17. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas. (2018b). Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589. https://arxiv.org/abs/1806.03589.

18. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

19. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.

20. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

21. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

22. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

23. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

24. Pathak, Deepak; Krahenbuhl, Philipp; Donahue, Jeff; Darrell, Trevor; Efros, Alexei A.. (2016). Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379. https://arxiv.org/abs/1604.07379.

25. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.

26. Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

27. Pearson, Karl. (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. https://doi.org/10.1080/14786440109462720.

28. Ulyanov, Dmitry; Vedaldi, Andrea; Lempitsky, Victor. (2017). Deep Image Prior. arXiv:1711.10925. https://arxiv.org/abs/1711.10925.

29. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

30. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.