Optimal prediction in the linearly transformed spiked model

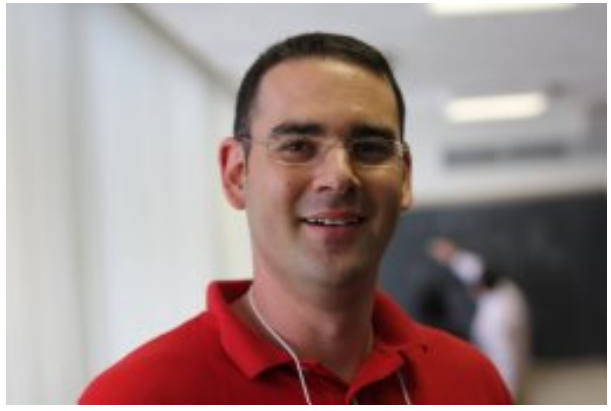
Edgar Dobriban

University of Pennsylvania

July 31, 2017

Collaborators

(a) William Leeb.
Princeton PACM



(b) Amit Singer.
Princeton Math+PACM

arxiv.org/abs/1611.10333, to be updated very soon

Linearly transformed spiked model

- Model $Y_i = A_i X_i + \varepsilon_i$, $i = 1, \ldots, n$
- Observe
  - $Y_i \in \mathbb{R}^{q_i}$, $A_i \in \mathbb{R}^{q_i \times p}$
  - Y_i: linear transform of unobserved signal of interest $X_i \in \mathbb{R}^p$
- Goal: recover X_i
  - "spiked" signal, lies on unknown low dimensional space
- Information loss:
  - Sampling noise ε_i
  - Transformation A_i: $q_i < p$
- Commutative case: $A_i^\top A_i$ diagonal

Examples

I Standard spiked model

II Cryo-Electron Microscopy (Cryo-EM)

III Missing data

Example I: Standard spiked model

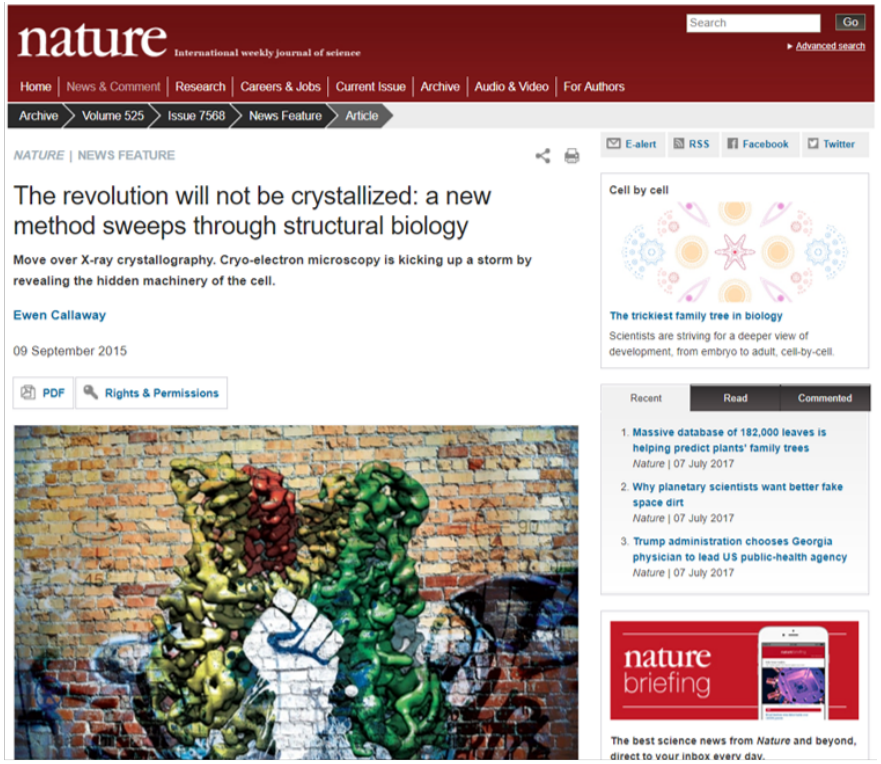
- $Y_i = X_i + \varepsilon_i$, $i = 1, \ldots, n$, so $A_i = I_p$
  - $X_i \in \mathbb{R}^p$ signal, lies on unknown low dimensional space
  - $\varepsilon_i \in \mathbb{R}^p$ noise
- Well studied under high-dimensional asymptotics: e.g., Johnstone (2001); Baik et al. (2005); Baik and Silverstein (2006); Paul (2007); Nadakuditi and Edelman (2008); Nadler (2008); Bai and Ding (2012); Bai and Yao (2012); Benaych-Georges and Nadakuditi (2012); Onatski (2012); Onatski et al. (2013, 2014); Nadakuditi (2014); Gavish and Donoho (2014); Johnstone and Onatski (2015); Hachem et al. (2015); Yao et al. (2015); Donoho et al. (2017).

Example II: Cryo-EM

- Mapping the structure of molecules without crystallizing them
- Imaging of heterogeneous samples, with mixtures of molecules or multiple conformations

Cryo-EM


Science
AAAS
Home
News
Journals
Topics
Careers
SHARE
PERSPECTIVE | BIOCHEMISTRY
The Resolution Revolution
Werner Kühlbrandt
Article
Figures & Data
Info & Metrics
eLetters
PDF
You are currently viewing the summary.
View Full Text
Summary
Precise knowledge of the structure of macromolecules in the cell is essential for understanding how they function. Structures of large macromolecules can now be obtained at near-atomic resolution by averaging thousands of electron microscope images recorded before radiation damage accumulates. This is what Amunts et al. have done in their research article on page 1485 of this issue (1), reporting the structure of the large subunit of the mitochondrial ribosome at 3.2 Å resolution by electron cryo-microscopy (cryo-EM). Together with other recent high-resolution cryo-EM structures (2–4) (see the figure), this achievement heralds the beginning of a new era in molecular biology, where structures at near-atomic resolution are no longer the prerogative of x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.
Science
Vol 343, Issue 6178
28 March 2014
ARTICLE TOOLS
RELATED CONTENT
Structure of the Yeast Mitochondrial Large Ribosomal Subunit
SIMILAR ARTICLES IN:
PubMed
Google Scholar
CITED BY...


nature
International weekly journal of science
Home
News & Comment
Research
Careers & Jobs
Current Issue
Archive
Audio & Video
For Authors
Archive
Volume 525
Issue 7568
News Feature
Article
NATURE | NEWS FEATURE
The revolution will not be crystallized: a new method sweeps through structural biology
Move over X-ray crystallography. Cryo-electron microscopy is kicking up a storm by revealing the hidden machinery of the cell.
Ewen Callaway
09 September 2015
PDF
Rights & Permissions
Cell by cell
The trickiest family tree in biology
Scientists are striving for a deeper view of development, from embryo to adult, cell-by-cell.
Recent
Read
Commented
1. Massive database of 182,000 leaves is helping predict plants' family trees
Nature | 07 July 2017
2. Why planetary scientists want better fake space dirt
Nature | 07 July 2017
3. Trump administration chooses Georgia physician to lead US public-health agency
Nature | 07 July 2017
nature briefing
The best science news from Nature and beyond, direct to your inbox every day.


nature methods
Techniques for life scientists and chemists
Home
Current issue
Comment
Research
Archive
Authors & referees
About the journal
home ▸ archive ▸ issue ▸ editorial ▸ full text
NATURE METHODS | EDITORIAL
Method of the Year 2015
Nature Methods 13, 1 (2016) | doi:10.1038/nmeth.3730
Published online 30 December 2015
PDF
Citation
Rights & permissions
Article metrics
The end of 'blob-ology': single-particle cryo-electron microscopy (cryo-EM) is now being used to solve macromolecular structures at high resolution.
Subject terms: Cryoelectron microscopy
The three-dimensional structure of a protein or protein complex provides crucial insights into its biological function. As a structure-determination technique, cryo-EM has played second fiddle to the higher-resolution approaches of X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. This is rapidly changing, however, thanks to recent technical advances that now allow near-atomic-resolution structures to be solved using cryo-EM. The time is right to celebrate single-particle cryo-EM as our Method of the Year.
Editors' pick
Statistics for Biologists
A free nature.com resource
Visit the collection
Science jobs
Science events
natureevents directory
Protein Interactions and Networks
09 October 2017 — 17 October 2017
Hinxton, United Kingdom
Mechanisms of Eukaryotic Transcription
29 August 2017 — 02 September 2017
N/A, United States
5th Annual European Congress on Clinical & Translational Sciences (EUSTM-2017)


Science
AAAS
Home
News
Journals
Topics
Careers
SHARE
REPORT
The 3.8 Å resolution cryo-EM structure of Zika virus
Article
Figures & Data
Info & Metrics
eLetters
PDF
You are currently viewing the abstract.
View Full Text
Abstract
The recent rapid spread of Zika virus and its unexpected linkage to birth defects and an autoimmune-neurological syndrome has generated worldwide concern. Zika virus is a flavivirus like dengue, yellow fever and West Nile viruses. We present the 3.8Å resolution
ARTICLE TOOLS
SIMILAR ARTICLES IN:
PubMed
Google Scholar
CITED BY...
CITING ARTICLES IN:
Web of Science (91)
Scopus (109)
RELATED JOBS FROM SCIENCECAREERS
Biochemistry

Deconvolution in cryo-EM

- **Goal:** Estimate unfiltered projections X_i from

$$Y_i = A_i X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

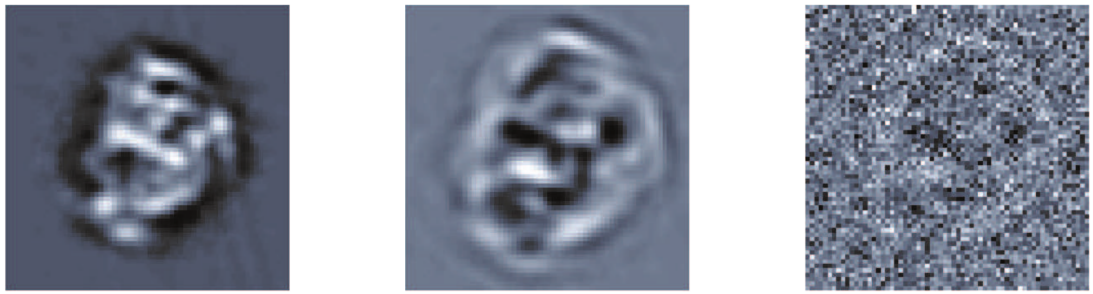


Figure : From left to right: X_i. $A_i X_i$. $A_i X_i + \varepsilon_i$

- A_i blurring convolution operator. Ill-conditioned. Diagonalized by Fourier Transform.

- Deconvolution arises in other imaging problems too

Other applications to cryo-EM

- Ab-initio 3D reconstruction
- Heterogeneity

Example III: Missing data

- $Y_i = A_i X_i + \varepsilon_i$
- A_i are coordinate-selection operators.
  - k-th row of A_i selects coordinate $l_i(k)$: $A_i(k, l) = \delta_{k, l_i(k)}$
  - $A_i^\top A_i$ diagonal, 1-s for observed entries, 0-s otherwise
- High-noise matrix completion.
  - Most prior work designed for low-noise scenarios. (Candès and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2009, 2010; Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Recht, 2011; Rohde et al., 2011; Jain et al., 2013)

Goals

- Develop general methods for observations of the form $Y_i = A_i X_i + \varepsilon_i$
- Task: Predict X_i.
- Desired properties:
  1. Applicable to big data.
  2. Weak distributional assumptions. (moments)
  3. Robust to high levels of noise.
  4. Statistical guarantees. (optimality)

Our approach/methodology

1. Random effects models: Best Linear Predictor (BLP, BLUP) (Searle et al., 2009)
2. Estimate unknown coefficients: Empirical BLP.
3. Equivalent to singular value shrinkage of a new of random matrix model.
4. RMT: Characterize spectrum.
5. "Normalize" to use optimal shrinkers (Nadakuditi, 2014; Gavish and Donoho, 2014).

1. Random effects models: Best Linear Predictor (BLP).

- "Spike" X_i: $X_i = \sum_{k=1}^{r} \ell_k^{1/2} z_{ik} u_k$, where $z_{ik} \dot{\sim} (0,1)$, $\ell_1 > \ldots > \ell_r > 0$, and $|u_k| = 1$.
- BLP $\hat{X}_i^B = LY_i$ minimizes $\mathbb{E}|\hat{X}_i^B - X_i|^2$
- $L = \mathrm{Cov}\,[X_i, Y_i]\,\mathrm{Cov}\,[Y_i, Y_i]^{-1} = \Sigma_X A_i^\top (A_i \Sigma_X A_i^\top + \Sigma_\varepsilon)^{-1}$
- Under conditions (see paper) show $\mathbb{E}|\hat{X}_i^B - \hat{X}_i^0|^2 \to 0$, where

$$\hat{X}_i^0 = \sum_{k=1}^{r} \tau_k u_k u_k^\top A_i^\top Y_i$$

for some $\tau_k > 0$.

2. Estimate unknown coefficients: Empirical BLP.

- BLP $\hat{X}_i^0 = \sum_{k=1}^{r} \tau_k u_k u_k^\top A_i^\top Y_i$
- u_k, τ_k parameters, must be estimated.
- $$\hat{X}_i = \sum_{k=1}^{r} \eta_k \hat{u}_k \hat{u}_k^\top A_i^\top Y_i$$

  where $\hat{u}_k$ are eigenvectors of cov mx of $A_i^\top Y_i$.
- Goal: find optimal η_k

3. Equivalence to singular value shrinkage.

- By exchangeability,

$$\mathbb{E}|\hat{X}_i - X_i|^2 = \frac{1}{n}\mathbb{E}|\hat{X} - X|^2,$$

  where $X = (X_1, \ldots, X_n)^\top$, $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_n)^\top$.
- Backprojected data $B = (A_1^\top Y_1, \ldots, A_n^\top Y_n)^\top = \sum_{k=1}^m \sigma_k \cdot \hat{u}_k \hat{v}_k^\top$,

$$\hat{X} = \sum_{k=1}^r \eta_k \sigma_k \cdot \hat{u}_k \hat{v}_k^\top$$

  is singular value shrinkage estimate.
- Thus, asy optimal η_k are determined by asy spectrum of B (e.g. Nadakuditi, 2014; Gavish and Donoho, 2014).

4. RMT: Characterize spectrum.

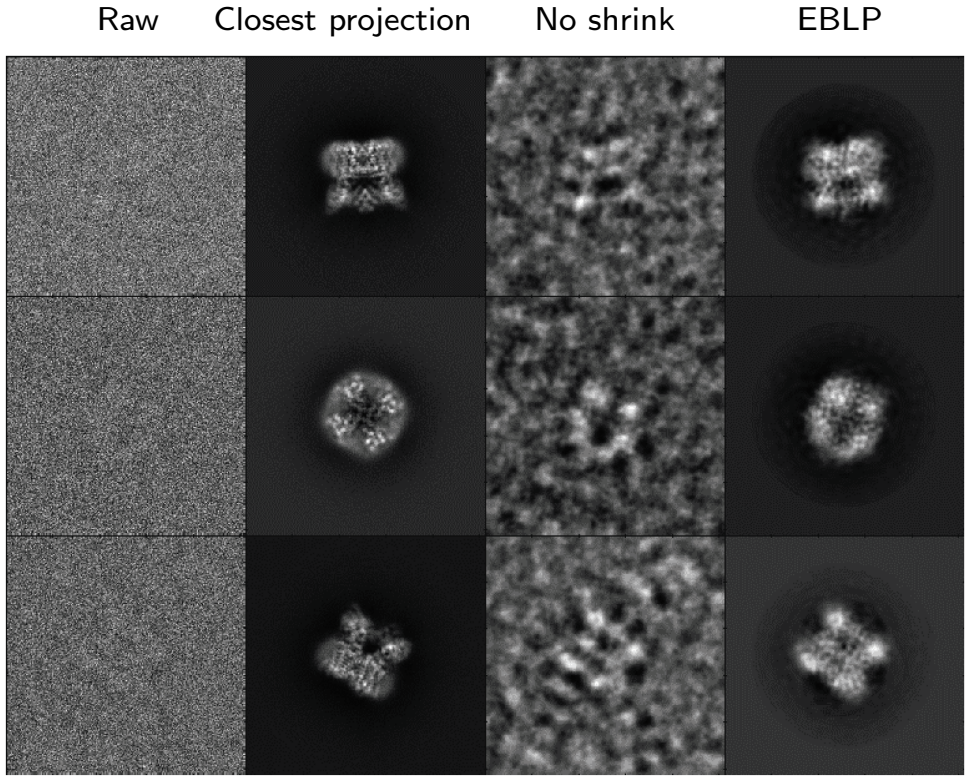
- Remains to find asy spectrum of $B = (A_1^\top Y_1, \ldots, A_n^\top Y_n)^\top$
- Non-standard random matrix model.
- $A_i^\top Y_i = A_i^\top A_i X_i + A_i^\top \varepsilon_i$.
  - Dependence between signal and noise.
  - General moment assumptions on X_i, A_i, ε_i, no invariance.
- Follow approach of Benaych-Georges and Nadakuditi (2012); extend "deterministic equivalents" tools of Bai et al. (2007).

5. "Normalize".

- Almost done... But optimal shrinkage coefficients η_k depend on $|\mathbb{E}[A_i^\top A_i]u_k|^2$, not clear how to estimate consistently.
- Normalize: $\hat{M}^{-1/2} A_i^\top Y_i$, where $\hat{M} = n^{-1} \sum_{i=1}^n A_i^\top A_i$. Everything goes through.

Results: Experimental data - TRPV1

Bhamre, Zhang, Singer (*Journal of Structural Biology, 2016*)

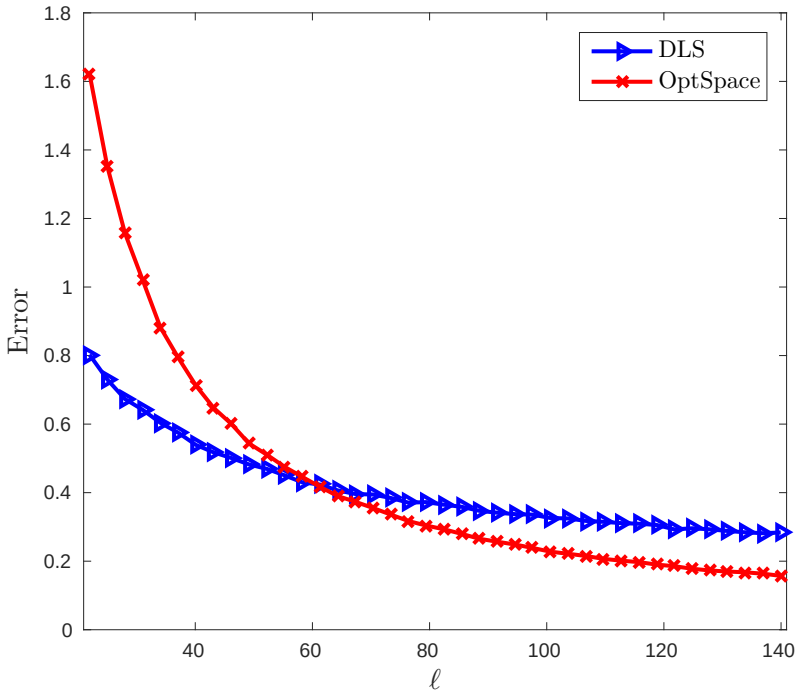


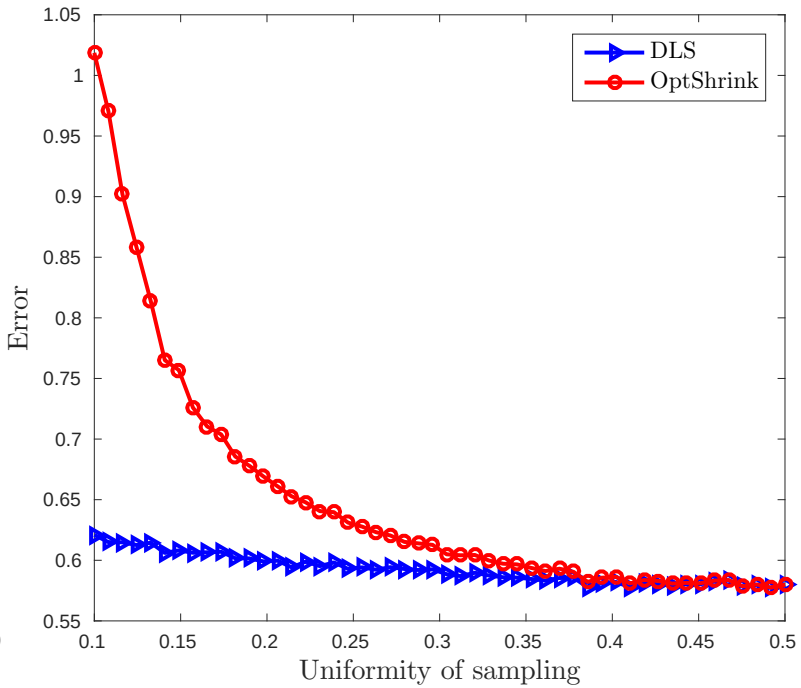
- TRPV1, K2 direct electron detector
- 35645 motion corrected, picked particle images of 256×256 pixels belonging to 935 defocus groups (Liao et al., Nature 2013)

Results: matrix completion

- We test our shrinkage method to other scalable matrix completion methods.

D, Leeb, Singer (*arXiv 2016*)





- Left: Comparison with OptSpace (Keshavan et al., 2009) for different SNRs. Our method outperforms in the high-noise regime.
- Right: Comparison with OptShrink (Nadakuditi, 2014) for different sampling uniformity levels. Our method outperforms OptShrink when the data is unevenly sampled across the rows of the matrix.

Conclusion

- Linearly transformed spiked model: $Y_i = A_i X_i + \varepsilon_i$, $i = 1, \ldots, n$
- Broad applicability: Image processing, Cryo-EM, Missing data
- EBLP Methods for predicting X_i: fast, weak assumptions, robust to noise, guarantees
- Key technical result: spectrum of backprojected data.

slides available at github.com/dobriban

Bai, Z. and Ding, X. (2012). Estimation of spiked eigenvalues in spiked models. *Random Matrices: Theory and Applications*, 1(02):1150011.

Bai, Z., Miao, B., and Pan, G. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532–1572.

Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177.

Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408.

Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Donoho, D. L., Gavish, M., and Johnstone, I. M. (2017). Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851, to appear in AoS*.

Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, 60(8):5040–5053.

Hachem, W., Hardy, A., and Najim, J. (2015). A survey on the eigenvalues local behavior

of large complex correlated wishart matrices. *ESAIM: Proceedings and Surveys*, 51:150–174.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.

Johnstone, I. M. and Onatski, A. (2015). Testing in high-dimensional spiked models. *arXiv preprint arXiv:1509.07269*.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998.

Keshavan, R. H., Oh, S., and Montanari, A. (2009). Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329.

Nadakuditi, R. R. (2014). Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018.

Nadakuditi, R. R. and Edelman, A. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *Signal Processing, IEEE Transactions on*, 56(7):2625–2638.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Onatski, A., Moreira, M. J., and Hallin, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231.

Onatski, A., Moreira, M. J., and Hallin, M. (2014). Signal detection in high dimension: The multispiked case. *The Annals of Statistics*, 42(1):225–254.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430.

Rohde, A., Tsybakov, A. B., et al. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Yao, J., Bai, Z., and Zheng, S. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press.