SVD vs OASIS: Mathematical Objectives Comparison

The Two Optimization Objectives

From Section S.3.E of the OASIS supplementary materials (Baharav, Tse, and Salzman):

SVD Objective

The Singular Value Decomposition (SVD) computes:

OASIS p-value Bound Objective

In contrast, OASIS optimizes the p-value bound by computing:

The Key Difference

The only difference is in how is normalized:

- **SVD uses:** $\sqrt{}$ (Euclidean/L2 norm)
- **OASIS uses:** (maximum/L-infinity norm)

Both use for the column vector .

What This Means

Notation

- is the centered and normalized contingency table
- is the row embedding vector (weights for rows/features)
- is the column embedding vector (weights for samples)
- is the test statistic (a bilinear form)

SVD Interpretation

- Maximizes the correlation-like quantity between rows and columns

- Penalizes based on its **total squared magnitude**
- Encourages to distribute weight across multiple rows
- **Effect:** Finds principal directions capturing the most variance
- This is equivalent to finding the largest singular value

OASIS Interpretation

- Maximizes the same numerator (test statistic)
- Penalizes based only on its **largest element**
- **Does not penalize** spreading weight across many rows (as long as max is controlled)
- **Effect:** Can put weight on many rows simultaneously
- This leads to the p-value bound via Hoeffding's inequality

Why OASIS Uses

The Statistical Reason

From the OASIS paper, the test statistic is:

Under the null hypothesis, are (approximately) independent random variables.

Hoeffding's inequality bounds the probability that a weighted sum of bounded random variables deviates from its mean. The bound depends on:

- The range of each random variable
- The **maximum weight** (not the sum of squared weights)

Key insight: For concentration inequalities, what matters is , not $\sqrt{}$.

The Mathematical Connection

Hoeffding's bound: If are independent with and is their weighted sum, then:

For OASIS's setup:

- Want to bound
- The bound involves
- With appropriate normalization, this becomes

Therefore: The term in the denominator is **exactly what appears in the p-value bound**, not an arbitrary choice.

Practical Implications

SVD Behavior

When you maximize :

Example: Suppose you have two rows with signal:

- Row 1 has moderate correlation with column pattern
- Row 2 has moderate correlation with column pattern

SVD solution: Put weight on both rows (e.g.,)

- Numerator: Gets contribution from both rows
- Denominator:
- **Combines signal from multiple rows**

OASIS Behavior

When you maximize :

Same example:

OASIS solution: Put weight on both rows (e.g.,)

- Numerator: Gets contribution from both rows (twice as much as SVD)
- Denominator:
- **Can freely combine multiple rows without additional penalty**

Key difference: OASIS is not penalized for using many rows (as long as no single weight is too large), while SVD is penalized by the sum of squares.

The Trade-off

SVD Advantages

- **Variance maximization:** Finds directions explaining most variation
- **Orthogonal decomposition:** Subsequent components are uncorrelated
- **Well-understood:** Decades of theory and practice
- **Stable:** Small data changes → small solution changes

OASIS Advantages

- **Statistical validity:** P-value bound is mathematically justified
- **Sparse-data robustness:** term handles low counts better
- **Interpretability (claimed):** Binary for clustering
- **Focused power:** Can concentrate on specific row patterns

SVD Disadvantages

- **No p-values:** Doesn't provide statistical significance
- **Asymmetric treatment:** Both rows and columns use L2 norm
- **Arbitrary scaling:** How "significant" is the first singular value?

OASIS Disadvantages

- **Computational complexity:** constraint makes optimization harder (NP-hard)
- **Local optima:** Alternating maximization doesn't guarantee global optimum
- **Data splitting:** Uses train/test split, throwing away half the data
- **Limited theory:** Asymptotic properties less developed than SVD

A Critical Perspective

Is This Really Better?

The claim: OASIS provides "statistically valid" p-values while SVD doesn't.

The reality:

1. **Both are fitting a model to high-dimensional data** (K rows, N columns)
2. **Both involve optimization** that may not find the global optimum
3. **Both make assumptions** (independence, distributional forms)
4. **OASIS adds data splitting** (train/test), which reduces power

The question: When K is in the thousands and N is in the tens:

- Is an "exact finite-sample p-value" more trustworthy than SVD's eigenvalue?
- Or have both methods entered the "complex model fit to sparse data" regime?

The Optimization Problem

Both objectives try to maximize , just with different normalization.

But: With K rows and N columns:

- There are possible binary vectors (if restricting to {-1, 1})
- Finding the optimal is **NP-hard** for both formulations
- **In practice:** Use greedy algorithms (alternating maximization)
- **Result:** You get a local optimum, not the global optimum

Implication: The "exact p-value" is for **the you found** , not the optimal . This introduces unquantified uncertainty.

When Does the Difference Matter?

Scenario 1: Strong Signal in Few Rows

Setup: 2-3 rows have strong signal, rest are noise

SVD: Puts most weight on those 2-3 rows, some weight on noise rows (due to L2 penalty)

OASIS: Puts weight primarily on those 2-3 rows (L∞ doesn't penalize this)

Winner: OASIS might have slightly better power (but probably small difference)

Scenario 2: Diffuse Signal Across Many Rows

Setup: 50 rows each have weak signal

SVD: Spreads weight across all 50 rows, captures cumulative signal

OASIS: Can also spread weight across 50 rows (no penalty from L∞ as long as weights are similar)

Winner: Similar performance, but SVD has cleaner interpretation

Scenario 3: Sparse Contingency Table

Setup: Many cells with zero or small counts

SVD: Standardization by $\sqrt{}$ can amplify noise in low-count rows

OASIS: Standardization by is more conservative

Winner: OASIS may be more robust (this is the claimed advantage)

The Bottom Line

Mathematical Perspective

The difference between and reflects:

- **SVD:** Variance decomposition framework
- **OASIS:** Concentration inequality framework

These are genuinely different mathematical approaches.

Statistical Perspective

The key question is: **Does the framework matter when K is large and N is small?**

Arguments for "No":

1. Both involve optimization that finds local optima
2. Both fit complex models to sparse data
3. Both make distributional assumptions that are approximate
4. OASIS throws away half the data (train/test split)
5. In high dimensions, the distinction between "exact p-value" and "useful approximation" becomes philosophical

Arguments for "Yes":

1. Having a p-value bound (even approximate) is better than no p-value
2. Concentration inequalities provide interpretable guarantees
3. Robustness to sparse data is valuable
4. The normalization is mathematically justified by Hoeffding

Practical Perspective

For most genomics applications:

- Prediction accuracy and biological validation matter more than exact p-values
- Replication in independent cohorts is the gold standard
- Whether you use SVD or OASIS is less important than:
  - Quality of data
  - Appropriate preprocessing
  - Biological interpretation
  - Independent validation

OASIS's real contribution may be:

- Not the p-value itself
- But rather: A different way to think about matrix decomposition
- That happens to be more robust to sparsity
- And provides a statistical framework (even if approximate)

Conclusion

The mathematical difference between SVD and OASIS is clear and principled:

This difference:

- **Is mathematically justified** (Hoeffding's inequality requires)
- **May provide practical benefits** (robustness to sparsity)
- **Enables p-value bounds** (unlike SVD)

But in the context of genomics with K >> N:

- **Both methods enter the "complex model" regime**
- **Both make approximations**
- **Both have limitations**
- **The "exact" p-value may be illusory** (due to optimization, data splitting, assumptions)

The deeper question remains: **In high-dimensional genomics, is optimizing for statistical validity (exact p-values) the right goal, or have we entered a regime where prediction, robustness, and biological interpretation matter more?**

References

1. Baharav, T.Z., Tse, D., and Salzman, J. (2024). "OASIS: An interpretable, finite-sample valid alternative to Pearson's X^2 for scientific discovery." *PNAS* 121(15). Supplementary Section S.3.E.
2. Hoeffding, W. (1963). "Probability inequalities for sums of bounded random variables." *Journal of the American Statistical Association* 58(301): 13-30.
3. Eckart, C. and Young, G. (1936). "The approximation of one matrix by another of lower rank." *Psychometrika* 1(3): 211-218. [Original SVD paper]