

Applications of the High-Dimensional Statistics to Astrophysics and Related Fields

Tsutomu T. Takeuchi, Yu Ogane, Kazuyoshi Yata, Kento Egashira,
Makoto Aoshima, Kohji Yoshikawa, Nanase Harada, Koichiro Nakanishi

October 2, 2026

Abstract

Modern spectral maps contain far more frequency channels than spatial samples. We present high-dimensional statistical methods for extracting physical information from such data, emphasizing ALMA observations of NGC 253. Noise-reduction and sparse principal component analysis identify line-intensity variations, rotation, and outflow signatures, while the larger ARCHEMI data set illustrates their use in detecting contaminated frequency ranges. We then introduce recent extensions to neutral hydrogen (HI) absorption forests, where physically important mean evolution need not produce a dominant covariance component. Optimal-transport coordinates separate total absorption from persistent line-of-sight morphology, providing a complementary route to interpreting high-dimensional spectra.

1 Introduction: physical information in spectral maps

Galaxies contain stars, interstellar gas and dust, and dark matter. Their evolution reflects interactions among these components across a wide range of spatial scales. Of particular importance is the conversion of interstellar gas into stars and the response of the surrounding gas to star formation. The interstellar medium includes ionized, neutral atomic, and molecular phases. Understanding how these phases are distributed and move within a galaxy is therefore essential to connecting the physics of molecular clouds with the evolution of the galaxy as a whole.

Spectroscopy provides access to these otherwise inaccessible physical conditions. Molecular transitions contain information about composition and excitation, while line positions and profiles reveal gas motions. A spectral map combines this information with spatial structure: every observed position has a spectrum, and every frequency channel provides a map. This is much richer than a single integrated spectrum, but it also creates a statistical difficulty. The diversity of spectral features can greatly exceed the number of spatial samples available to characterize it.

Let n denote the sample size and d the number of measured variables. In a spectral cube, a spatial sample supplies a vector of d frequency channels. The regime $n \ll d$ is called high-dimensional, low-sample-size, or HDLSS. Increasing spectral coverage or resolution can increase d substantially without a corresponding increase in n . Moreover, neighboring map pixels can be correlated, so the nominal number of spatial samples should not automatically be interpreted as the number of independent observations. The difficulty is not simply the storage of a large data matrix. Estimating its covariance structure requires statistical methods appropriate to the available sampling (Jung and Marron, 2009; Yata and Aoshima, 2012).

Our main observational example is the nuclear starburst of NGC 253. Its ALMA spectra are rich in molecular lines, and the spatially resolved data offer a direct test of whether high-dimensional methods can recover interpretable features without first reducing the spectra to a prescribed set of integrated line intensities (Ando et al., 2017; Takeuchi et al., 2024). We first explain the statistical representation, then follow the analysis from the original spectra through removal of systemic rotation to the identification of smaller-scale motions. The larger ARCHEMI working data set provides a further example in which component analysis first reveals problems in the observations themselves.

The final part introduces HI forest absorption spectra. These also have many channels and few sightlines, but their physical information is distributed differently between the mean and covariance. This contrast motivates the recent transport-geometric and mean-aware analyses developed in our absorption studies, referred to below as Papers I and II.

2 High-dimensional PCA and spectral interpretation

2.1 The data matrix and its dual covariance

For principal component analysis (PCA), write the spectrum at sample a as $\mathbf{x}_a \in \mathbb{R}^d$, and collect the spectra as columns of X . Subtracting the channel-by-channel sample mean gives $\tilde{X}$. The sample covariance and its dual are

$$\hat{S} = \frac{\tilde{X}\tilde{X}^\top}{n-1}, \quad \hat{S}^D = \frac{\tilde{X}^\top\tilde{X}}{n-1}. \quad (1)$$

Although $\hat{S}$ is $d \times d$, the dual is only $n \times n$. They have the same nonzero eigenvalues, and centering bounds their rank by $n-1$. The dual formulation therefore permits computation in the much smaller sample space while retaining the nonzero sample-covariance structure. It does not, however, create additional samples or eliminate statistical estimation error.

In classical PCA, the leading eigenvectors describe directions of large sample variance. In the HDLSS regime, contributions from many dimensions can bias the eigenvalues and alter the estimated directions. Noise-reduction PCA, abbreviated NRPCA, addresses this high-dimensional contribution under specified population conditions (Yata and Aoshima, 2012). For centered data, the eigenvalue correction used in our analysis is

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\text{Tr}(\hat{S}^D) - \sum_{s=1}^j \hat{\lambda}_s}{n-j-1}, \quad 1 \leq j \leq n-2. \quad (2)$$

The subtraction estimates an average contribution from the remaining sample eigenvalues. This is a correction to covariance estimation, not a license to regard every weak spectral feature as instrumental noise. Its usefulness must be assessed in the data under study.

2.2 Eigenspectra, sparse loadings, and score maps

A principal direction has one coefficient per frequency channel and can therefore be plotted as an eigenspectrum. Projecting each spatial spectrum onto that direction produces a score, which can be returned to its original position on the map. Eigenspectra and score maps answer complementary questions: which spectral features vary together, and where does that variation occur?

Automatic sparse PCA, or A-SPCA, makes the first question more explicit by retaining the channels that carry the important loadings and setting other coefficients to zero. Its application to NGC 253 is described by Takeuchi et al. (2024). Several selected channels may belong to the same molecular line, so a component supported on a small number of channels can reflect only a few underlying spectral features. The sparse representation is useful for identifying those features; NRPCA retains the fuller eigenspectrum needed to describe the spectral variation beyond the selected channels.

The interpretation requires both representations. A loading concentrated on the core of a bright line may track intensity, while a loading with different signs on opposite sides of a line may indicate a displacement or an asymmetric profile. The score map then tests whether that spectral behavior follows a galactic-scale gradient or a localized structure. The sign of a principal component is conventional: reversing both its loading and its scores leaves the represented variation unchanged. A negative score is not, by itself, evidence for approaching gas.

A further distinction concerns centering. PCA describes deviations from the sample mean, not the mean itself. A variation common to all samples can therefore be absent from the centered covariance even when it is physically important. The NGC 253 analysis below demonstrates the effectiveness of covariance-based interpretation, while the absorption example in Sec. 4 shows why it is sometimes necessary to retain the mean as a separate statistical object.

3 NGC 253: from molecular spectra to gas dynamics

3.1 A line-rich starburst observed with ALMA

NGC 253 provides a favorable observational setting because its nuclear region contains intense star formation, rich molecular emission, and complex gas motions. The Band 7 observations of Ando et al. (2017) resolve diverse star-forming structures on approximately 10 pc scales. The beam corresponds to about 8×5 pc, and the spectral windows sample the 340–365 GHz range. The representative spectra and continuum map in Fig. 1 illustrate how strongly the line spectra vary among the central structures.

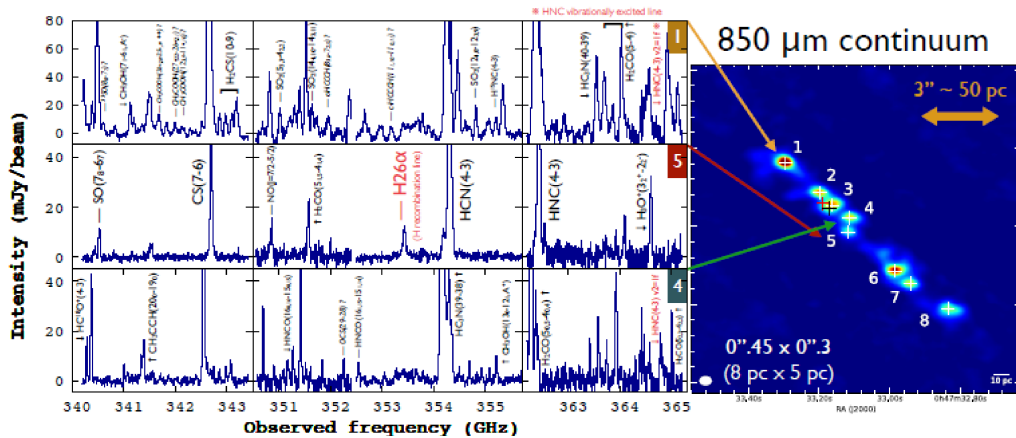


Figure 1: Representative ALMA Band 7 spectra and the 850 μ m continuum map of the central region of NGC 253. The molecular spectra differ between the labeled structures even within this compact region (Ando et al., 2017).

The working data matrix contains $n = 231$ spatial spectra and initially $d = 2248$

channels. The objective is not merely to identify individual transitions, but to determine which combinations of spectral features organize the spatial variation. Reducing the data to a short list of integrated lines before analysis could conceal information in line wings or small profile distortions. High-dimensional PCA instead allows those frequency-resolved features to contribute to the statistical decomposition (Takeuchi et al., 2024).

3.2 What the uncorrected spectra reveal

The initial analysis deliberately retained the Doppler shifts across the galaxy. The first two high-dimensional principal components accounted for most of the prominent global organization: PC1 approximately traced overall line intensity, whereas PC2 described the systematic spectral displacement produced by rotation. In the PC1–PC2 score diagram, the latter appeared as a characteristic butterfly-like distribution. Mapping the scores back onto the sky made the kinematic interpretation much clearer than the score diagram alone.

Figure 2 compares the original PC2 map with the velocity field measured from HCN(4–3). Both show a coherent gradient along the observed region. This is an important validation: the analysis did not use a velocity label to define PC2, yet the resulting coordinate follows the large-scale motion. Deviations from this correspondence identify positions at which the common rotation pattern does not fully explain the spectra.

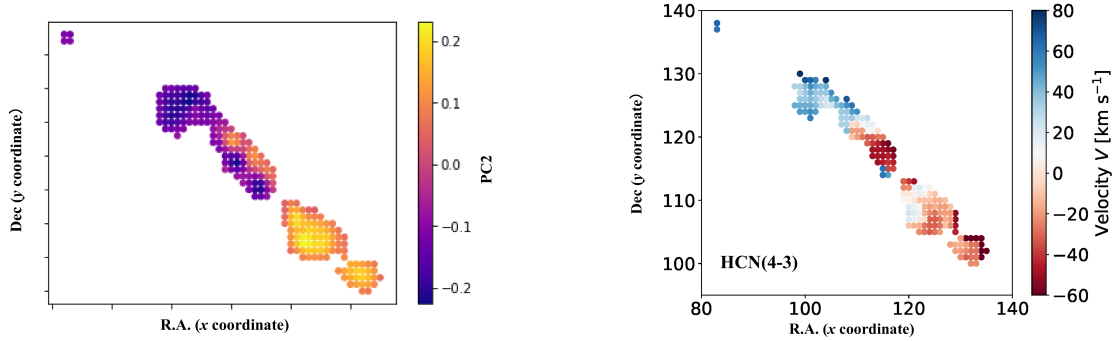


Figure 2: Left: PC2 score map from the original spectra. Right: the HCN(4–3) velocity field. Their spatial correspondence identifies the dominant kinematic contribution to PC2. The two color bars represent different quantities and should not be compared numerically (Takeuchi et al., 2024).

A-SPCA also localized the spectral channels responsible for these components. The dominant features included HCN(4–3) and HNC(4–3). Only a small subset of the original channels was needed to describe the important loadings, and several selected channels traced different parts of the same transition. Thus, spatial organization and spectral localization supported the same interpretation: total line strength and large-scale Doppler displacement were the leading sources of variation in the uncorrected cube.

3.3 Removing rotation without discarding line profiles

The next step was to remove the common rotational displacement so that subtler differences could become visible. We estimated the local velocity using HCN(4–3), HNC(4–3), and CS(7–6), averaged the estimates, and shifted the spectra to a common kinematic frame. The common spectral coverage after this operation contains $d = 1971$ channels. This correction aligns line positions; it does not replace each line by its integrated intensity or discard the residual width and asymmetry.

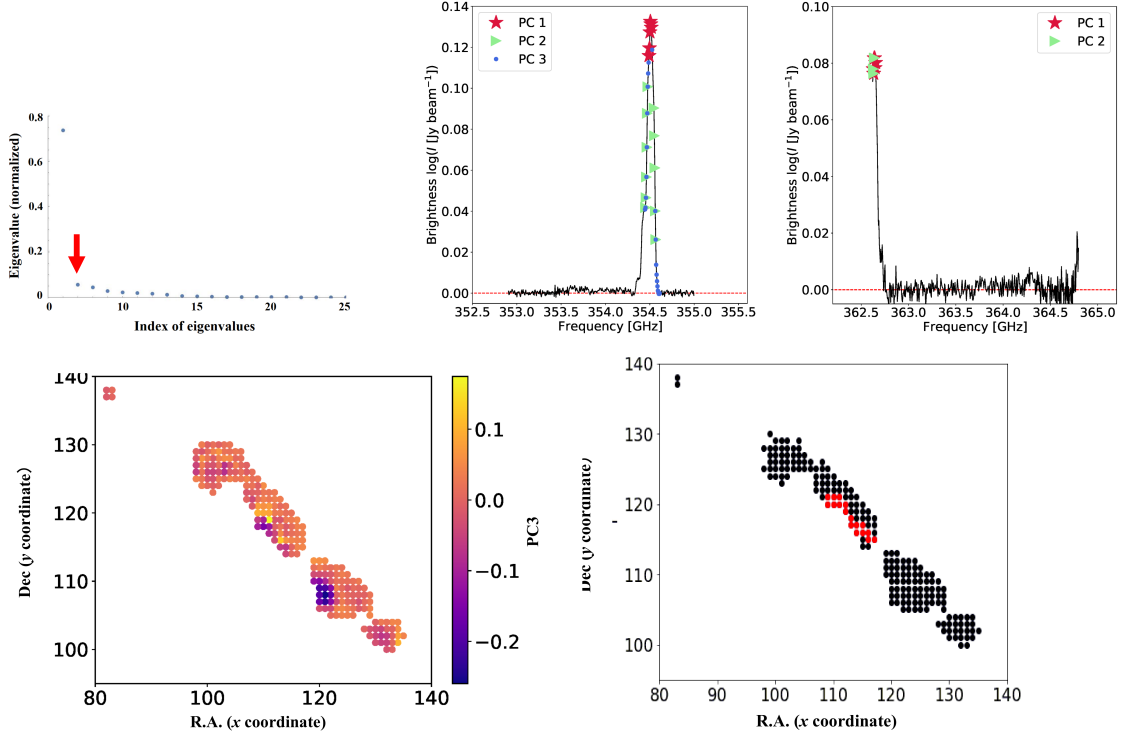


Figure 3: Doppler-corrected analysis. Upper left: the eigenvalue distribution. Upper right: channels selected by sparse PCA around HCN and HNC. Lower left: PC3 score map. Lower right: anomalous positions in the velocity field, marked in red. The spatial and spectral evidence supports an outflow interpretation for the localized higher-component signal (Takeuchi et al., 2024).

The butterfly-like score pattern disappeared after correction. PC1 became more strongly dominant in the eigenvalue distribution, while the higher eigenspectra emphasized smaller spectral differences. Importantly, component numbers refer to the particular data representation: PC2 after correction is not the same statistical direction as PC2 before correction. The change in the eigenspectra and score maps is precisely what allows the large-scale rotational contribution to be separated from local gas motions.

3.4 Local expansion and the molecular outflow

The corrected PC1 again approximately represents total line intensity. The corrected PC2 emphasizes line wings, including those of HCN, and is associated with a localized expansion on a scale of roughly 10 pc. PC3 and partly PC4 describe blueshifted profile distortions, whereas another higher component is associated with redward distortions. These associations are inferred jointly from the eigenspectra and their spatial score patterns, rather than from the signs of the scores alone (Takeuchi et al., 2024).

The region with negative PC3 in the displayed sign convention lies near the root of the known southwest molecular streamer. Its correspondence with the anomalous velocity region provides a second spatial diagnostic (Fig. 3). The earlier analysis compared this location with molecular and optical observations of the outflow and found consistent spatial evidence. The high-dimensional decomposition therefore recovers an astrophysically recognizable structure from the full spectral map.

The physical result is the identification and localization of coherent velocity-dependent spectral variation. It is not, by itself, a measurement of the outflow mass, kinetic energy, or mass-loss rate. Those quantities require additional physical modeling. The value of the statistical analysis is that it selects the relevant spectral features and positions before such detailed modeling, while distinguishing them from the much stronger variations associated with line intensity and rotation.

3.5 ARCHEMI: quality control before further physical interpretation

The ARCHEMI working cube extends this approach to approximately 85–163 GHz, with $d = 50702$ channels and of order $n \lesssim 2000$ spatial samples. The final spatial sample can be smaller after noisy positions are assessed. Relative to the Band 7 example, the working data have lower spatial resolution and higher spectral resolution. The broader ALCHEMI survey provides the molecular-inventory context (Martín et al., 2021). The result reported here concerns quality control of the full spectral working cube, rather than a new physical decomposition.

The first result was not a new gas component, but a data-quality diagnosis. Several initial principal-component maps showed stripes, nearly featureless gradients, or patterns inconsistent with the source morphology. Inspection of their contributing frequency ranges identified strong atmospheric absorption. Removing the affected ranges produced component maps with coherent structure concentrated on the central molecular region (Fig. 4).

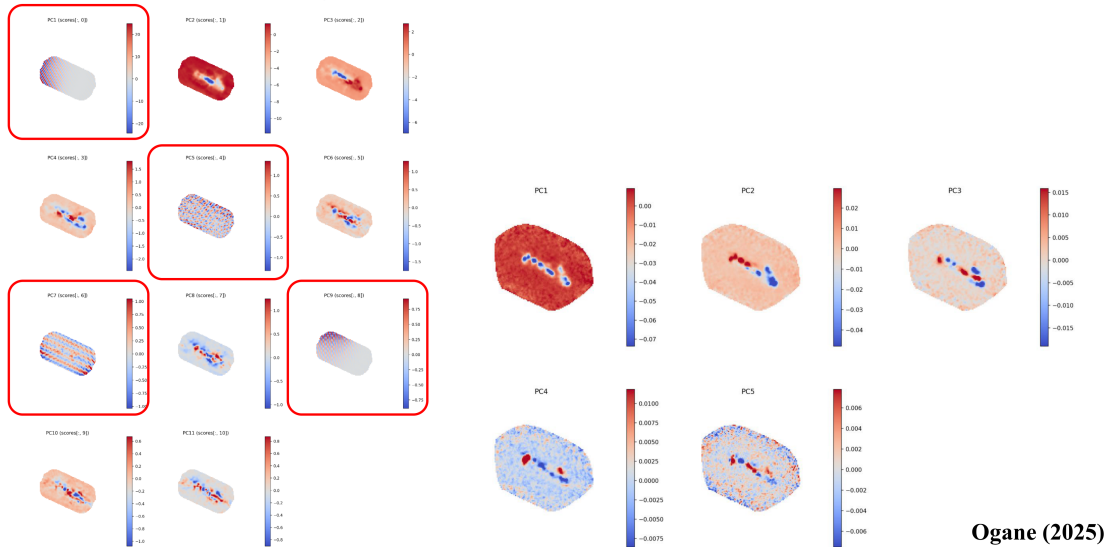


Figure 4: ARCHEMI quality control reported by Ogane in the 2025 presentation. Left: the initial component maps, with the original red outlines highlighting anomalous patterns. Right: all five displayed component maps after removal of the damaged frequency ranges. The affected frequency ranges were identified as suffering from strong atmospheric absorption (Ogane et al. 2026 in progress).

This example extends the interpretation of an eigenspectrum. The channels with large loadings can identify not only an astrophysical transition or a line wing, but also a frequency interval whose observational quality compromises the analysis. A coherent artifact can carry substantial covariance and become a leading component. Its statistical prominence does not make it a physical structure in the galaxy.

The result documented at this stage is the detection and removal of these contaminated intervals. The cleaned maps provide the starting point for further molecular and kinematic interpretation, but no additional physical decomposition of the cleaned cube is claimed here. Together with the Band 7 study, this shows why spectral loadings and spatial score maps should be examined jointly throughout a high-dimensional analysis.

4 HI absorption line forests on radio quasar spectra

4.1 Observation geometry and physical identifiability

HI 21-cm forest observations probe foreground neutral gas through absorption against background radio sources (Furlanetto and Loeb, 2002). Many spectral channels can be available along relatively few sightlines. Unlike an emission map, however, a forest spectrum is not a direct three-dimensional gas map: density, spin temperature, velocity, and line broadening enter a common observation operator. Paper I separates degeneracies of this operator from information discarded by the subsequent statistical representation.

For a normalized local line profile ϕ , the optical depth and ideal absorption depth are schematically

$$\tau_a(v) = \int q_a(\chi) \phi_a(v - \mathcal{V}_a(\chi); \chi) \, d\chi, \quad g_a = 1 - \mathcal{R}_a[e^{-\tau_a}], \quad (3)$$

where q_a is the absorption coefficient, $\mathcal{V}_a$ includes expansion and peculiar velocity, and the fixed instrumental response $\mathcal{R}_a$ preserves positivity and the continuum. Instrumental smoothing acts after the exponential attenuation. For a nonnegative profile of positive total absorption, define

$$M_g = \int g(v) \, dv, \quad p_g(v) = g(v)/M_g, \quad (4)$$

and let $Q_g(q)$ be the generalized inverse of the cumulative distribution of p_g . A distance retaining both total absorption and its spectral arrangement is

$$d_\beta^2(g, h) = \beta \left(\log \frac{M_g}{M_h} \right)^2 + \frac{1}{v_*^2} \int_0^1 [Q_g(q) - Q_h(q)]^2 \, dq. \quad (5)$$

Here $\beta > 0$ and $v_* > 0$ are fixed scales. The second term is the one-dimensional squared Wasserstein distance (Amari and Matsuda, 2022). The transported measure describes absorption, not gas-particle trajectories, and M_g is not the gas mass.

In regular profile families, the zero directions of the pullback metric coincide with the kernel of the linearized observation map. This supplies a criterion for distinguishing an unobservable physical change from a change lost through finite features, local aggregation, or compression. Controlled examples in Paper I recover a column-density-to-spin-temperature ratio, a temperature under stated broadening assumptions, and parameters of gas ensembles. They do not establish universal superiority over conventional statistics or a correctly specified flux likelihood. The central result is a separation of what the observation identifies from what a chosen analysis retains.

The physical examples clarify this distinction. An isolated 21 cm component directly constrains a column-density-to-spin-temperature ratio, not the two quantities separately. A kinetic temperature can be inferred from its width only after the nonthermal broadening is specified. For ensembles, the controlled calculations estimate a clumping parameter of the adopted density-generating law and a transverse correlation length from several

sightlines. These are model parameters, not a reconstruction of every unresolved gas structure. In particular, when the probability law along a single sightline is held fixed, the transverse scale is identified only through the dependence between different sightlines. Point-estimation accuracy and independently calibrated uncertainty regions must consequently be assessed under the same physical and observational assumptions.

4.2 Mean evolution need not be a covariance spike

Paper II applies the HDLSS viewpoint to an archived TNG50-derived mock (Nelson et al., 2019). Three snapshots near $z \simeq 10, 9$, and 8 contain the same 12 sightline labels, with 1024 spectral bins per epoch. Concatenating the three epochs gives 3072 features per sightline. The validation groups are the 12 matched sightlines, not 36 independent spectra, and all belong to one periodic simulation volume.

The integrated optical depth is dominated by epoch evolution: redshift accounts for 98.48% of its matched two-way variance. At the pixel level, however, the between-epoch scatter B accounts for only 7.05% of the total scatter $T = B + W$, where W is the within-epoch scatter. Nevertheless, 95.54% of the trace of B belongs to its first eigenvalue. The mean evolution is therefore nearly rank one but low trace. The much larger within-epoch scatter has a first-eigenvalue fraction of only 13.33%. These are finite-catalog diagnostics, not a determination of a population asymptotic spiked model.

Let $\mathbf{m}_{\text{lin}}$ denote the mean profile change along a centered linear redshift contrast. The within-epoch sample covariance has rank 33 in 1024 dimensions, and its entire nonzero eigenspace captures only 2.02% of the squared norm of $\mathbf{m}_{\text{lin}}$ (Fig. 5). The remaining signal is outside the empirical covariance span, not necessarily outside the population covariance support. The comparison is also conditional on the chosen representation. Aligning the epochs permits their mean profiles to be compared in one spectral space; concatenating epochs and centering across sightlines removes their shared evolution from the covariance target. Neither operation makes the epoch dependence physically unimportant. The mismatch instead shows why a covariance-only search can fail to represent a known, coherent mean change. To retain the mean explicitly, we estimate

$$\widehat{W}_{\alpha,\gamma} = (1 - \alpha)\widehat{W} + \alpha \text{diag}(\widehat{W}) + \gamma \bar{w} I_p, \quad (6)$$

$$\widehat{\mathbf{v}}_{\alpha,\gamma} \propto \widehat{W}_{\alpha,\gamma}^{-1} \widehat{\mathbf{m}}_{\text{lin}}, \quad \bar{w} = \text{Tr}(\widehat{W})/p, \quad (7)$$

with $p = 1024$, identity matrix I_p , $0 \leq \alpha \leq 1$, and $\gamma > 0$. Nested sightline-wise validation holds out all epochs of one sightline together. The resulting score correctly orders the epochs on every held-out sightline. Its redshift RMSE, 0.1138, is comparable to 0.1069 from log integrated optical depth. This diagnoses recovery of the three known epoch states, not an observational redshift measurement. The high-dimensional direction is predominantly a positive, broad-band weighting of absorption.

4.3 Total absorption and persistent spectral shape

The transport coordinates make this distinction explicit. We use $g = 1 - e^{-\tau}$ and retain M_g in physical velocity units. For shape comparison, each epoch's velocity interval is mapped to $u \in [0, 1]$. This separates the interval width from the relative arrangement of absorption within it. Comparing unnormalized physical-velocity quantiles would otherwise include the systematic dilation of the different spectral windows.

Redshift accounts for 98.50% of the variance of $\log M_g$, whereas sightline identity accounts for 90.30% of normalized-quantile variation. After removing the quantile centroid,

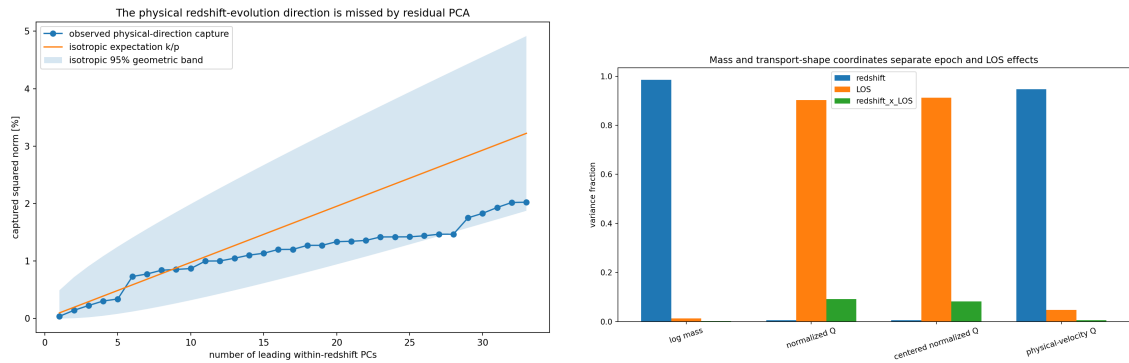


Figure 5: Simulated HI forest results. Left: the linear epoch-mean direction is largely outside the empirical within-epoch PCA span. Right: total absorption mainly records epoch evolution, whereas normalized quantiles mainly record sightline identity. Physical-velocity quantiles retain the additional effect of different spectral-window spans. The decomposition is conditioned on the matched mock catalog.

nearest-neighbor matching identifies every sightline across each of the three epoch pairs, in both directions. Mass alone does not retain this identity. Conversely, normalized shape alone has little epoch-predictive power in this catalog.

Approximately 70% of the shape variation remains after projecting out location and one scale-like direction, and 91.54% of this residual is a sightline main effect. Persistent morphology is therefore not merely a centroid or width difference. Diffusion maps (Coifman and Lafon, 2006) give a stable first shape coordinate, but its correlation with shape PC1 is 0.9983; the 12 sightlines do not demonstrate an additional nonlinear manifold. These findings are conditioned on the single-volume mock and its fixed forward model. They identify complementary absorption coordinates, not a unique reconstruction of the underlying three-dimensional gas.

5 Brief concluding remarks

High-dimensional analysis of NGC 253 extracts interpretable intensity and velocity structures from line-rich spectra, while the ARCHEMI example demonstrates its value for data-quality diagnosis. The absorption studies add a complementary lesson: a physical trend can reside in the mean rather than in a dominant covariance component. Keeping total absorption and spectral shape separate allows epoch evolution and persistent sightline morphology to be studied without conflating their statistical roles.

References

- Amari, S., and Matsuda, T. 2022, *Annals of the Institute of Statistical Mathematics*, 74, 33–47.
- Ando, R., Nakanishi, K., Kohno, K., et al. 2017, *ApJ*, 849, 81.
- Coifman, R. R., and Lafon, S. 2006, *Applied and Computational Harmonic Analysis*, 21, 5–30.
- Furlanetto, S. R., and Loeb, A. 2002, *ApJ*, 579, 1.

- Jung, S., and Marron, J. S. 2009, *Annals of Statistics*, 37, 4104–4130.
- Martín, S., Mangum, J. G., Harada, N., et al. 2021, *A&A*, 656, A46.
- Nelson, D., Pillepich, A., Springel, V., et al. 2019, *MNRAS*, 490, 3234.
- Takeuchi, T. T., Yata, K., Egashira, K., et al. 2024, *ApJS*, 271, 44.
- Yata, K., and Aoshima, M. 2012, *Journal of Multivariate Analysis*, 105, 193–215.