Principal Component Analysis of Type Ia Supernova Spectra: Improving Standardisation and Twinning

Reis Tyson
Supervisor: Dr Mathew Smith
Lancaster University Physics Department

27 April 2026

Abstract

Type Ia supernovae are powerful cosmological distance probes used to constrain the expansion history of the Universe. Standardisation techniques such as SALT2 aim to reduce residual scatter in distance estimates, but this reduction is insufficient for precision cosmology. We propose the inclusion of supernova spectra, decomposed through Principal Component Analysis in standardisation techniques. Here, we outline a unified framework to extend the traditional SALT2 standardisation and 'twin' physically similar supernovae with spectroscopic information. Our defined PCA distance used to describe spectral similarity shows a strong correlation ($r = 0.97$) with full RMS comparisons. Hence, this framework provides both global and local standardisation procedures within a unified PCA-based spectral representation. Our SALT2 extension achieves scatter reductions of 6% while maintaining a large dataset and the closest twins show very low residuals (0.03-0.05 mag). We also provide an efficient twin screening procedure in PCA-space intended to be applied prior to existing twinning techniques. The methodology outlined shows standardisation improvement, it is deployable to large datasets and provides a potential basis for significant scatter reduction and high redshift supernova evaluation, proposed to enable tighter cosmological constraints.

Contribution Statement

The Type Ia supernova dataset used in this work was obtained from the public ZTF SNIa DR2 release. The light-curve data, including fitted SALT2 parameters (e.g. x_1, c, x_0) and redshift values, were provided as part of this dataset, along with associated spectroscopic files containing wavelength and flux measurements.

All data handling and analysis was performed by the author. This includes sample selection, spectral preprocessing (restframe correction interpolation and normalisation), and construction of the analysis ready datasets.

The implementation of principal component analysis, including decomposition, reconstruction, and coefficient extraction, was carried out independently. The extension of SALT2 standardisation using PCA-derived spectral components, the definition of a PCA-based spectral distance metric, and the development and evaluation of spectral twinning and twin-screening methodologies were conducted by the author.

All statistical analysis and interpretation of results alongside production of figures were performed by the author.

Contents

1 Introduction

1.1 Cosmology from Distance Measures

Observational cosmology aims to constrain the Universe's expansion history through observables, particularly those that measure cosmic distance-redshift relations. Measurements of these distance indicators allow for diagnostic plots such as Hubble diagrams [1], providing constraints on cosmological parameters such as H_0, Ω_m, and the dark energy equation of state. This framework additionally enables evaluation of theoretical models, via luminosity distance $D_L(z)$ [2], and constrains the expansion history of the Universe. However, this methodology is fundamentally underpinned by the requirement for both a sufficient sample size and observable accuracy. One such candidate for distance indicators is Type Ia Supernovae.

1.2 Type Ia Supernova as Standard Candles

Type Ia Supernovae, referred to as SNIa throughout this report, are valuable distance indicators in cosmology. SNIa have high intrinsic luminosity $M_V \approx -19.3$ [3], allowing observations across a wide redshift range. SNIa also have relatively uniform peak luminosity due to similar progenitor systems [4]. Ultimately, SNIa are powerful cosmic distance probes, used to discover the accelerated expansion of the Universe [5] [6].

Comparing the intrinsic luminosity and observed flux of SNIa, assuming an inverse square law, distance can be inferred. This distance is expressed through $D_L(z)$ accounting for the expansion of the Universe. Models can then be fit and cosmology inferred through Hubble diagrams.

Several observational surveys, ZTF (used for this study) [7], DES [8], Pan-STARRS [9], provide large SNIa sample sizes spanning large redshift ranges. This increasing sample size reduces statistical uncertainty in cosmological measurements. Therefore, systematic error in measurements dominates uncertainty. The accuracy of each SNIa measurement is critical and improving accuracy is a priority.

The major source of systematic error in supernova cosmology arises through the underlying standard candle assumption. Although SNIa have consistent luminosities, SNIa are not intrinsically identical. Observed luminosity exhibits both intrinsic variation and environmental dependencies. Therefore, SNIa cannot be used as true standard candles without correction (standardisation), and are so called "standardisable" candles. Empirical standardisation methods aim to reduce the scatter in inferred luminosity, correcting for observable correlations. The effectiveness of these standardisations directly impacts the precision of cosmological constraints.

1.3 Standardisation & Residual Scatter

The SALT2 methodology [10] is the typical SNIa standardisation approach and is used as a baseline throughout this investigation. The method applies a global light curve standardisation model across the SNIa dataset. An empirical model is used, constructed with the light curve parameters stretch and color (x_1, c). This approach does not separate intrinsic and environmental effects and compresses SNIa physics, resulting in unmodeled variation between SNIa contributing to residual scatter. Applying SALT2 standardisations alone significantly reduces scatter, achieving $0.15\,mag$ [11] in well controlled samples. However, this scatter does persist in large datasets highlighting the dominance of systematic standardisation error over statistical uncertainty. This residual scatter is the dominant limitation in extracting precise cosmological results from SNIa observations.

Alternate approaches aim to improve standardisation beyond SALT2, with techniques incorporating SNIa spectral information. Pairwise matching or "twinning" [12] of SNIa also

provides a novel approach applying local standardisations between pairs or subsets of SNIa.

The persistent scatter indicates that photometric, light curve based standardisation is insufficient, motivating the inclusion of additional parameters such as spectroscopic information.

1.4 Spectral Information & Twins Approaches

While typical approaches use limited observables and do not fully capture physical variation between SNIa, spectral analysis of SNIa provides potential to further reduce standardisation scatter. Spectra encode SNIa composition, temperature and ejecta velocity, which have been shown to correlate with the intrinsic luminosity of SNIa [13]. Therefore, spectra can provide a more complete physical description than light curves alone. Prior standardisation studies using spectroscopic data have shown further scatter reductions, for example, SNEMO achieving $0.10\,mag$ [14].

Spectroscopic twinning is a local approach to standardisation involving pairwise comparisons of SNIa, where "twins" are SNIa pairs with highly similar spectra. This methodology assumes similar spectra correlate to similar intrinsic luminosities, enabling a relative standardisation between SNIa.

However, spectroscopic standardisation methods require high signal to noise spectra, introduce significant computational cost that scales with spectral dimensionality and are often challenging to scale to large datasets. Despite these limitations, spectral methods consistently reduce scatter highlighting that capturing more physical information produces better standardised SNIa.

Principal component analysis provides a method to reduce the dimensionality of SNIa spectra producing a compact spectral representation, which is central to the methodology outlined in this report. This enables the efficient use of spectra across large datasets. Standardisation through deconstructed spectra is therefore a natural enhancement to current standardisation approaches in order to reduce residual scatter and apply to precision cosmology.

1.5 Project Aims

We intend to improve SNIa standardisation incorporating spectroscopic information efficiently via principal component analysis, addressing practicality concerns. The project aims are as follows:

- Evaluate principal component analysis as a method to represent SNIa spectra, analysing the effectiveness of dimensional reduction, spectral reconstruction accuracy and the variance captured by principal components.
- Extend SALT2 standardisation with principal component analysis derived spectral terms. This is evaluated through the impact of this addition on Hubble diagram scatter.
- Define a spectral distance metric, quantifying spectral similarity through decomposed spectra.
- Identify twin SNIa using the spectral distance metric, evaluating whether twin pairs matched in this space show reduced luminosity scatter.
- Develop a twin candidate screening procedure to efficiently identify potential twins across large datasets. The success of this method is evaluated through screening accuracy with ROC used to assess screening performance.

Fundamentally, this study aims to provide and test the validity of a unified framework for global and local SNIa standardisations (SALT2 extension and twinning, respectively). Spectral information is captured efficiently, permitting analysis with a large number of SNIa. If

successful, these methods show potential for higher precision cosmology through improved distance.

2 Background Theory

2.1 Physics of Type Ia Supernova

Supernovae are catastrophic stellar explosions, with distinct classes triggered by a variety of physical mechanisms and progenitors. SNIa, a specific supernova population, are characterised by their standardisable peak luminosity, and are a key observable used widely in cosmology.

SNIa originate in binary star systems, in which a white dwarf accretes mass from a companion. This accretion increases the mass of the white dwarf, approaching the Chandrasekhar Limit [15]. Near this mass ($1.4M_{\odot}$ for a non-rotating white dwarf [3]), density and temperature increase significantly favouring carbon fusion. These degenerate conditions prevent pressure-temperature regulation, facilitating thermonuclear runaway, resulting in a catastrophic stellar explosion. Since the mass at explosion spans a narrow range, the intrinsic peak luminosity of SNIa is relatively uniform.

During the explosion of the white dwarf, a large amount of unstable Nickel-56 is produced, triggering the decay chain [16]:

$$^{56}\mathrm{Ni} \rightarrow {}^{56}\mathrm{Co} \rightarrow {}^{56}\mathrm{Fe} \tag{1}$$

The mass of ^{56}Ni produced determines the peak luminosity and the decay chain regulates the light curve evolution, with ^{56}Co decay affecting the late time decline rate [17], therefore the light curve tail.

Although this theory predicts a narrow range of intrinsic SNIa luminosities, differing ^{56}Ni mass results in varying luminosities and observations show both variation in SNIa luminosity and spectral features. Environmental effects, such as dust [18] and host galaxy properties [19], also significantly impact SNIa observations. Therefore, SNIa cannot be used as standard candles and are as such "standardisable candles".

It is important to note that while the SNIa theory discussed is the leading progenitor theory, there is increasing research favouring multiple models [20]. While not fully understood, multiple scenarios reinforce the need for standardisation and can explain the diversity of SNIa and introduce sub-classes of SNIa. The methodology outlined in this work shows potential to differentiate these classes for standardisation; however, this is not the focus of the report. Ultimately, the assumption that all SNIa have consistent intrinsic luminosity is not strictly valid. In order to reduce scatter, standardisation is essential enabling SNIa to be used as distance probes in cosmology.

2.2 Supernova Spectra

A spectrum is a plot of flux as a function of wavelength, tracking the distribution of emitted light across a wavelength range. Spectra are widely taken in astrophysics, obtained through spectroscopic measurements.

As shown in figure 1 SNIa spectra consist of broad features with a general lack of sharp absorption lines. This is due to Doppler broadening [21] as a distribution of ejecta velocities results in varying Doppler shifts for each absorbing particle, creating broadened absorption lines. These features encode SNIa characteristic properties such as ejecta velocity, composition and structure.

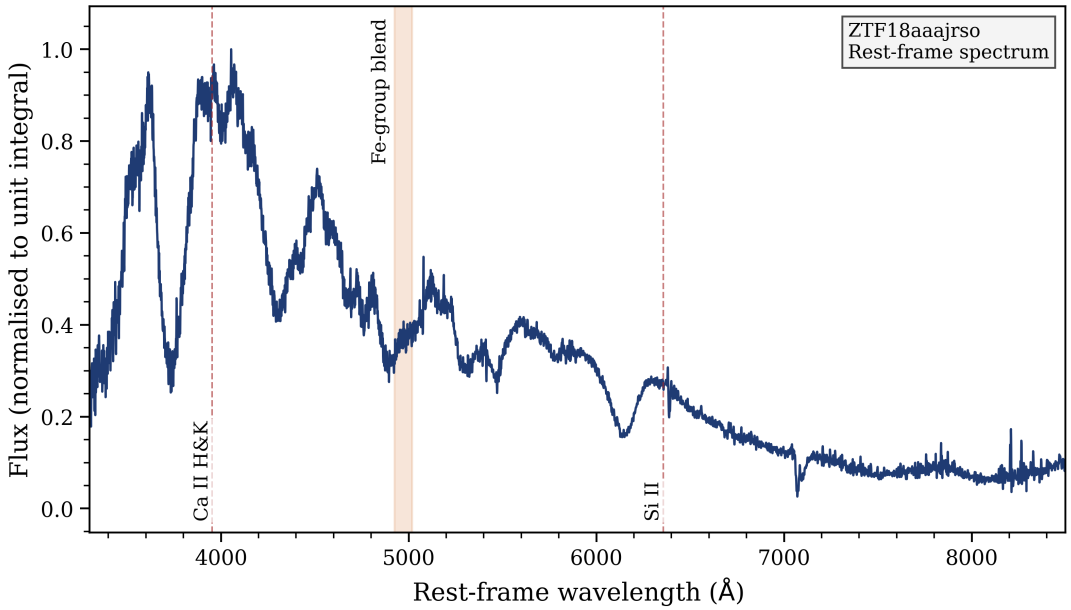

Figure 1: Rest-frame spectrum for SNIa ZTF18aaajrso. The flux is normalised to unit integral and rescaled to unit peak for clarity. Several approximate rest-frame wavelength values for absorption features are overlaid such as the Ca II H&K doublet ($\lambda\lambda\,3934,\ 3968\,\mathring{A}$) and prominent Si II absorption ($\lambda\,6355\,\mathring{A}$) [22]. Notice the shift in absorption features relative to their rest-frame wavelengths, most noticeably the Si II absorption dip positioned left of the overlaid line, as high ejecta velocities cause blue-shifting of absorption features. The shaded region highlights an Fe II dominated absorption blend ($\lambda\lambda\,4924,\ 5018\,\mathring{A}$) [23]. The spectra shown encode SNIa composition, structure and ejecta velocity, containing high-dimensional SNIa characteristic information, motivating the use of spectral information in standardisation.

Absorption features arise from specific elements, providing insight into SNIa composition, with Si and Ca particularly prominent. SNIa also contain iron-group elements (Fe, Co), and all spectra show an apparent lack of hydrogen, an observational defining feature of SNIa. These features allow ejecta velocity to be determined. With high photospheric velocity ($\sim$ $10{,}000\,\mathrm{km\,s^{-1}}$ [23]), the absorption minima are blue-shifted relative to rest-frame wavelengths. The blueshift of the Si II dip left of the illustrated wavelength is particularly prominent in figure 1.

However, features are not typically isolated, such as the Fe-group absorption blend shown in figure 1, transitions overlap and elements produce multiple features. Spectra are an evolving, high-dimensional description of SNIa characteristics, providing significantly more detailed physical information than light curves. Spectral variation has been shown to correlate with luminosity [24], motivating the use of spectroscopic information in SNIa standardisation to reduce scatter and facilitate precision cosmology. This poses significant challenges as spectra are complex, high-dimensional SNIa descriptors and so are difficult to compare directly or use in models without introducing significant computational costs. A compact representation of SNIa spectra through dimensional reduction can provide an efficient and applicable set of spectroscopic information, primed for large datasets.

2.3 Principal Component Analysis

Principal component analysis (PCA) [25] is a valuable tool used to simplify high dimensional information such as SNIa spectra. Spectra consist of values across thousands of wavelength bins, but since much of the physics acts across multiple wavelengths, not all are required. PCA identifies the dominant sources of variation allowing spectra to be described as a small number of principal components (PCs). A given spectral dataset can be represented as

$$X \in \mathbf{R}^{N \times M} \tag{2}$$

where N is the number of spectra and M the number of wavelength bins, and each spectrum can be described as

$$\mathbf{x_j} = (f_{\lambda_1}, f_{\lambda_2}, ...f_{\lambda_M}) \tag{3}$$

Therefore, the mean spectrum in the dataset is

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x_j}, \tag{4}$$

and the mean centred spectra are

$$\mathbf{x_j}' = \mathbf{x_j} - \overline{\mathbf{x}} \tag{5}$$

The PCA is then applied to the mean centred spectra, capturing variation relative to the mean spectrum. PCA determines maximal variance through

$$C\mathbf{e_i} = \lambda_i \mathbf{e_i} \tag{6}$$

where C is the covariance matrix, λ_i are the eigenvalues and $\mathbf{e_i}$ the eigenvectors. The eigenvectors correspond to the principal components and the eigenvalues indicate the variance encoded in each corresponding principal component such that

$$\lambda_1 > \lambda_2 > ... > \lambda_M. \tag{7}$$

Each spectrum can then be represented in PCA space through the eigenvectors. PCA coefficients a_{ij} encode how strongly each PC contributes to the spectrum given by

$$a_{ij} = \mathbf{x_j}' \cdot \mathbf{e_i}. \tag{8}$$

The spectra in PCA space

$$\mathbf{x_j} = \overline{\mathbf{x}} + \sum_{i=1}^{K} a_{ij} \mathbf{e_i} \tag{9}$$

encode the spectral diversity in a reduced K number of coefficient values rather than M wavelength bins. Since PCs are ordered in descending variance often only the first few components are required to capture the majority of the variance allowing for $K \ll M$.

2.4 Spectral Distance Metrics

Since spectra can be described as vectors in M dimensional space (equation 3), the difference between two spectra $\mathbf{x_A}$ and $\mathbf{x_B}$ is given by the spectral distance

$$D_{AB} = \|\mathbf{x_A} - \mathbf{x_B}\| = \sqrt{\sum_{m=1}^{M} \left(f_{\lambda_m}^{(A)} - f_{\lambda_m}^{(B)}\right)^2} \tag{10}$$

In PCA space this distance can be determined as

$$D_{AB} = \sqrt{\sum_{i=1}^{K} \left(a_{iA} - a_{iB}\right)^2} \tag{11}$$

the Euclidean distance between PC coefficients a for the K retained PCs. As PCA represents the spectra with a reduced set of values, distance measured in PCA space can act as a compact measure of spectral difference. In general, weighted distance can be defined as

$$D_{AB} = \sqrt{\sum_{i=1}^{K} w_i \left(a_{iA} - a_{iB}\right)^2} \tag{12}$$

where the chosen weighting w_i reflects the underlying assumptions of PC importance in describing spectral variation. For example, a variance weighting ($w_i = \lambda_i$) emphasises PCs capturing more variance while an inverse variance weighting ($w_i = \lambda_i^{-1}$) down-weights higher variance capturing PCs, resulting in a Mahalanobis[26]-like distance metric.

In this work, an unweighted distance is adopted 3.4.2 providing a simple spectral similarity measure. Given that PCs are ordered by variance and PC1 accounts for significant variance, a variance weighting produces a distance dominated by PC1, reducing sensitivity to spectral features. An inverse weighting causes a distance metric dominated by low PCs, amplifying noise and minor features, which may reduce the physical interpretability of the distance metric. An unweighted metric provides a simple balance between the dominant structure and secondary features. In general if variance weighting is required, higher order PCs can be manually down-weighted to prevent them dominating distance.

2.5 Cosmological Distance

On large scales, spacetime is not Euclidean, due to the expansion of the Universe. Distance is therefore redshift dependent. Luminosity distance [2] is calculated as a function of redshift, z,

$$D_L(z) = (1+z)\frac{c}{H_0}\int_0^z \frac{dz'}{E(z')} \tag{13}$$

where c is the speed of light, H_0 the Hubble constant and $E(z)$ defined as

$$E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda} \tag{14}$$

for flat ΛCDM cosmology, where Ω_m and Ω_Λ are the matter and dark energy density parameters. Cosmological theory and observations can be compared through distance modulus μ defined as

$$\mu = m - M = 5 \log_{10} \left(\frac{D_L}{10\,pc} \right) \tag{15}$$

where SNIa observations provide apparent magnitude, m, values and M is the absolute magnitude of SNIa, empirically calibrated and refined through standardisation. Construction of a Hubble diagram, of $\mu(z)$, allows cosmological models can be constrained. Standardisation is essential to determine accurate cosmic distances with SNIa and the accuracy of μ depends on standardisation quality, reducing scatter in these cosmic measurements facilitates precision cosmology with tighter constraints on cosmological parameters.

3 Methodology

The constants and values used throughout this work are shown in table 1 below.

Table 1: Key parameters and assumptions used throughout this analysis. Cosmology is implemented using Astropy [27] [28].

Quantity	Symbol	Value / Description	Type
Cosmology			
Hubble constant	H_0	70 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$	Fixed
Matter density parameter	Ω_m	0.3	Fixed
Dark energy density parameter	Ω_Λ	0.7 (flat Universe)	Derived
Cosmology model	–	Flat ΛCDM (Astropy)	Fixed
Spectral Processing			
Wavelength range	–	3600–8200 Å	Fixed
Interpolation step	–	5 Å	Fixed
Phase selection	–	±2 days around peak	Fixed
Normalisation	–	Unit integral	Fixed
PCA Configuration			
Number of components	K^*	15	Fixed
Sigma clipping	–	e.g. 3σ (varied)	Variable
Standardisation			
Stretch parameter	x_1	SALT2 parameter	Fitted
Colour parameter	c	SALT2 parameter	Fitted
Absolute magnitude	M	Empirically calibrated	Fitted
SALT2 coefficients	α, β	Light-curve correction terms	Fitted
Twinning / Distance Metric			
Distance metric	D	Euclidean (unweighted)	Fixed
PCs used in distance	–*	First 15 PCs	Fixed

* Unless otherwise stated, all PCA analyses use $K = 15$ principal components.

The method outlined explores a PCA-based approach to Type Ia supernova standardisation inspired by SNEMO [14] PCA. Standardisation approaches take two forms, identifying spectral twins [12] and SALT2 [10] extensions using principal component coefficients. Figure 2 below illustrates the workflow of the methods.

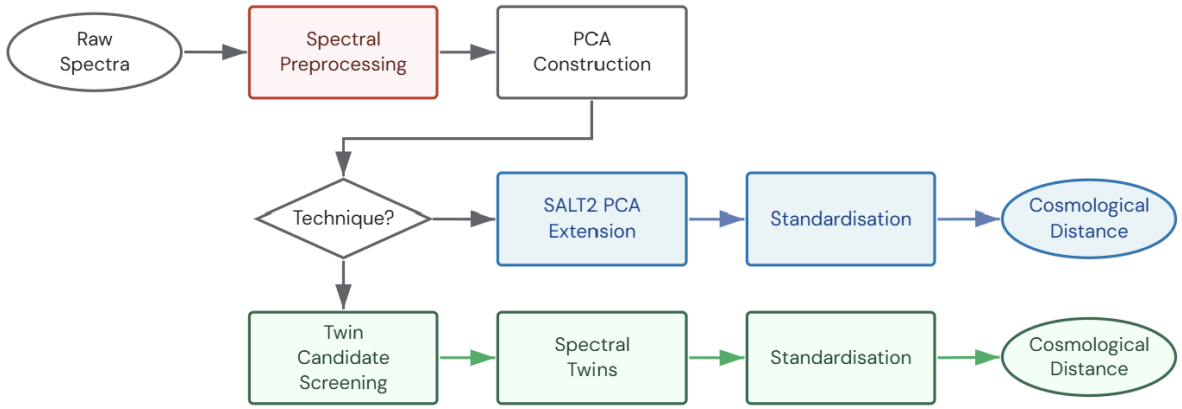

Figure 2: Flowchart illustrating the procedure to standardise raw Type Ia supernova spectra with two processes, a SALT2 extension for Hubble diagram standardisation (blue) and a spectral "twinning" (green).

3.1 Data Sample

The SNIa data used throughout this work are sourced from the public ZTF SNIa DR2 data release [11]. The data takes two forms, the first lightcurve data for 3629 SNIa, with file name `snia_data.csv`. From this file, the flux, x_0, stretch, x_1, and color, c, parameters are extracted for analysis. Second, spectroscopic data containing wavelength and corresponding flux values (some of which contain flux errors), are obtained from a spectra archive. The two can be matched through ztfname, providing both lightcurve and spectroscopic descriptions of the SNIa.

A sample is then constructed from this dataset, selecting only SNIa with a spectrum taken within 2 days of peak luminosity. For SNIa with multiple spectra files within this 2 day phase range, only one spectrum is assigned.

Spectra are drawn from a heterogeneous set of instruments, with SEDM providing the majority of observations and instruments such as the Liverpool Telescope, P200, NTT, and Keck providing a significant number of SNIa spectra. (See appendix A for a full list of instruments and file naming convention). In order to maximise the sample, spectroscopic error is neglected in this investigation to provide a consistent sample format. An extension to the work discussed in this report is to consider errors in the methodology.

After implementing the phase and spectral selection criteria, a final sample of 920 SNIa is obtained for preprocessing and analysis.

3.2 Spectra Preprocessing

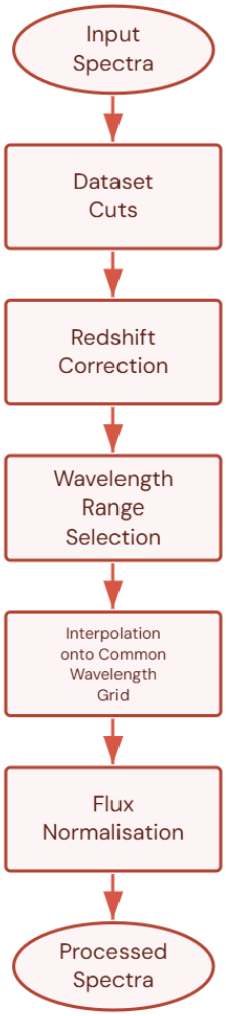

Figure 3: Flowchart illustrating the procedure to prepare Type Ia supernova spectra for PCA decomposition.

Before PCA can be undertaken several adjustments must be made to the spectra to ensure a consistent form. First a number of cuts were made to the dataset, so that the procedure can be used to generate a SALT2 extension, the typical SALT2 dataset cuts [11] were applied limiting

$$\text{stretch}, \quad x_1 \in [-3, 3],$$
$$\text{color}, \quad c \in [-0.2, 0.8],$$
$$\text{redshift}, \quad z < 0.06.$$

The phase of the spectra was also factored and the dataset is filtered to only spectra within 2 days of peak luminosity.

Once the cuts are made, next the spectra must be shifted into the rest frame correcting for redshift. For a wavelength point λ_{obs} in a supernova spectra, the rest frame wavelength is given by

$$\lambda_0 = \frac{\lambda_{obs}}{1+z} \tag{16}$$

where z is the redshift of the supernova. To allow PCA on the spectra, they must all cover the same wavelength range, therefore a wavelength cut is made to the data. This cut was taken at $3600 - 8200$Å in order to maximise the number of spectra remaining and ensure a sufficient wavelength range remains. Any spectra with maximum or minimum values inside this range were cut from the dataset. Of the 920 initial cut spectra, 787 remain after this wavelength cut.

PCA requires all of the spectra to be in the same format, therefore a common wavelength grid is used to provide a consistent set of wavelength values. Linear interpolation was used to achieve this since it is a simple approach, well suited to the small gaps between wavelength values, ensuring that artificial variation is not generated and the overall spectral shape remains post interpolation. Between two neighbouring points (λ_1, F_1) and (λ_2, F_2) for an intermediary wavelength λ its flux was calculated as

$$F(\lambda) = F_1 + \frac{F_2 - F_1}{\lambda_2 - \lambda_1}(\lambda - \lambda_1). \tag{17}$$

This allowed all spectra to be evaluated over a common wavelength grid from $3600-8200$Å with spacing $\Delta\lambda = 5$Å. Finally, before the spectra can be decomposed through PCA, the relative flux variation must be compared rather than absolute changes; therefore, all flux values were normalised through integral normalisation. The normalised flux, $\hat{f}$, is defined as

$$\hat{f} = \frac{f(\lambda)}{\int_{\lambda_{min}}^{\lambda_{max}} f(\lambda) d\lambda}. \tag{18}$$

This processing stage provided an interpolated grid of normalised flux values over a $3600-8200$Å range with steps of $\Delta\lambda = 5$Å suitable for PCA implementation for 787 spectra.

3.3 PCA Decomposition & Reconstruction

Once the spectra are processed into a uniform format, PCA can be conducted, reducing the high dimensional spectra into a lower dimensional representation. This method takes the processed spectra and inputs them through a Scikit-Learn [29] PCA package, which first mean centres the data. This captures the dominant modes of variation between SNIa resulting in a dataset consisting of principal component (PC) coefficients for each SNIa spectrum. Each spectrum can then be reconstructed as a linear addition of eigenspectra, where each i^{th} spectrum

$$F_i(\lambda) = \overline{F}(\lambda) + \sum_{k=1}^{N_{pc}} X_{i,k} E_k \tag{19}$$

where $\overline{F}_i$ is the mean spectrum, $X_{i,k}$ the extracted PC coefficient value and $E_k(\lambda)$ the k^{th} eigenspectrum. The output coefficients $X_{i,k}$ relate to the spectrum's position in PCA space, therefore similar spectra (similar SNIa) have similar PC coefficients and sit in the same region in PCA space. These coefficients can then be used for standardisation and twins identification.

For this method, the first 15 PCs are retained capturing 93% of the datasets variance. Lower PCs represent the global spectral structure while higher PCs capture finer details, rare spectral features and noise. Figure 4 illustrates several eigenspectra corresponding to PCs generated from this PCA.

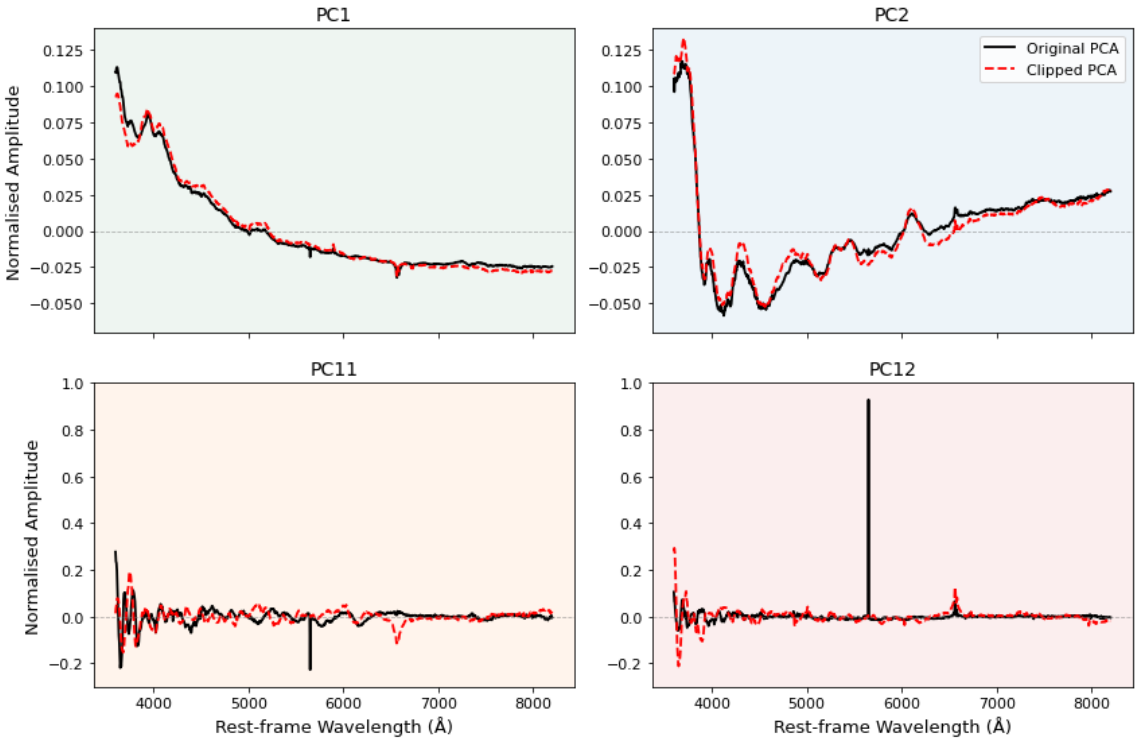

Figure 4: Illustration of principal components 1, 2, 11, 12 before and after 3σ range clipping.

The first eigenspectra, PC1, encapsulates the majority of the datasets spectral variation (65% after clipping) followed by PC2, PC3, sequentially. However, PCA is sensitive to extreme spectra. As shown in figure 4 components 11 and 12 have prominent spikes, this indicates that these eigenspectra are generated to represent specific SNIa with prominent spectral features. These eigenspectra offer a weak representation of the overall dataset, removing the outlier SNIa

that cause their construction and re-conducting PCA provides a more detailed description across the dataset. Figure 5 illustrates the dataset cut to improve the eigenspectra.

A clipping is therefore applied to each distribution of PC coefficients, cutting the range to 3σ range such that all SNIa are removed which satisfy

$$|X_{i,k}| > 3\sigma_k \tag{20}$$

for any k^{th} PC to maximise PCA description and maintain a reasonably sized dataset. The method relies on the assumption that all SNIa have reasonably similar intrinsic luminosity permitting standardisation, this cut must therefore be strict to ensure this assumption holds.

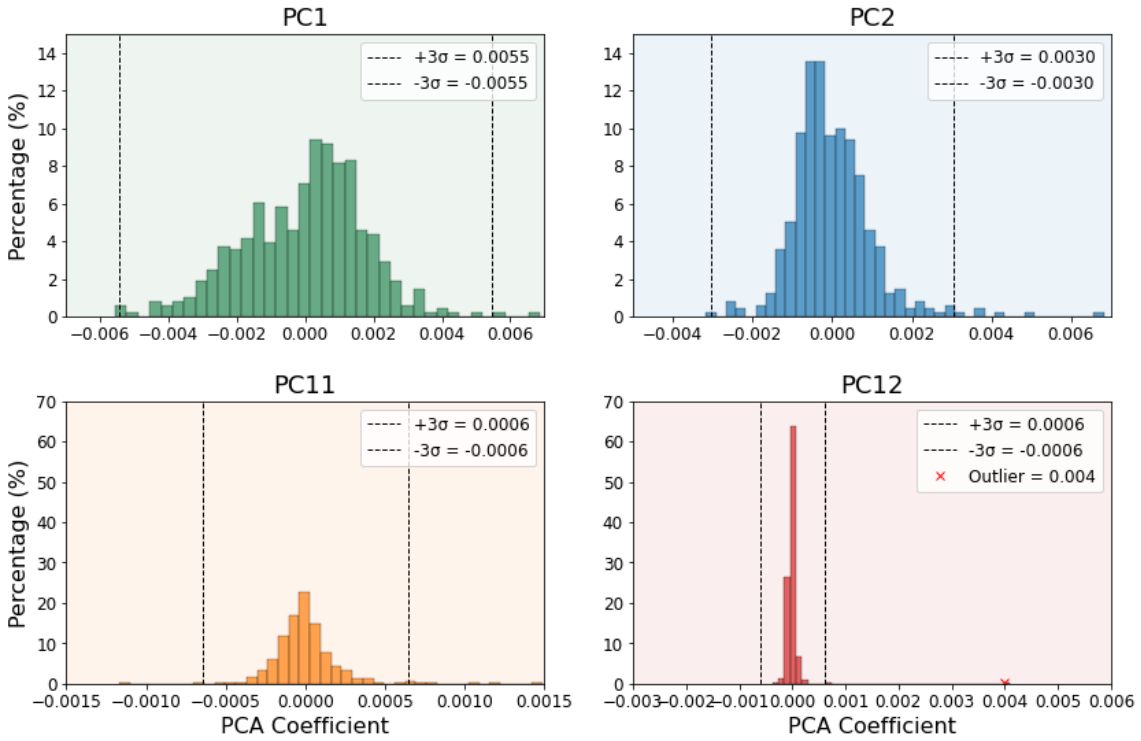

Figure 5: Histograms illustrating the distribution of the SNIa principal component coefficients for PC1, 2, 11, 12 shown in figure 4.

Figure 5 highlights that the majority of SNIa cluster in a narrow coefficient range. Outliers in this PC space correspond to strong spectral features and rare events. This results in certain PCs (PC11 and PC12) existing only to describe specific SNIa or subpopulations. The PC12 histogram (bottom right) shows an outlier with a coefficient value of 0.004, this outlier spectrum contains a prominent spectral feature aligning with the PC12 eigenspectrum spike. Therefore, the PC12 eigenspectrum exists solely to capture the variation in this one outlier spectrum. The PC11 coefficient distribution also contains SNIa outside of the 3σ range, again with spectral features related to the PC11 spike. Since these eigenspectra exist to map outlier SNIa outside of a 3σ range, implementing this cut and recomputing PCA generates more physically meaningful eigenspectra better representing typical SNIa. Crucially since these PC coefficients are used to compute a Euclidean distance in PC space, this cut ensures that this distance is representative of spectral diversity.

However, while for this method peculiar SNIa are removed, this pipeline highlights the potential for PCA to be used to identify unusual SNIa (or transients in general) or to identify SNIa sub-populations, though this report focusses on standardisation.

3.4 Supernova Standardisation Methods

In order to extract cosmology from Type Ia supernova (SNIa) observations, the scatter of standardised magnitudes must be minimised to reduce uncertainty in distance estimates. Traditional SALT2 standardisation corrects supernova luminosities using light-curve stretch (x_1) and colour (c), typically achieving Hubble residual scatters of order ~ 0.15 mag [30] in well-controlled samples. This introduces a standardisation in which the distance modulus is given by

$$\mu = m_B - M + \alpha x_1 - \beta c. \tag{21}$$

where m_B is the observed magnitude, M the SNIa's absolute magnitude and α and β are the standardisation coefficients for stretch and color respectively. However, this two parameter standardisation does not fully capture SNIa diversity. Improving standardisations to consider spectral diversity, therefore SNIa diversity, scatter can be reduced and distance measures improved.

PCA of SNIa spectra can be used in a number of ways for standardisation including extending simple SALT2 models to factor spectral diversity (3.4.1) and for "twinning" (3.5 & 3.4.2).

3.4.1 SALT2 PCA Extension

One approach to improve standardisation is to extend the simple SALT2 model beyond two light curve parameters. By introducing PCA coefficients to the SALT2 standardisation, spectral diversity is now factored into magnitude corrections with the distance modulus, μ, given by,

$$\mu = m_B - M + \alpha x_1 - \beta c + \sum_i \gamma_i p_i \tag{22}$$

where γ_i act as global coefficients fit across the dataset. In this work the first 15 PCs are used capturing 93% of the dataset's spectral variance. This model now factors both light curve parameters shown to impact the observed magnitude m_B [31] and spectral diversity encoding SNIa explosion physics correlated to SNIa luminosity [32].

3.4.2 Spectral Twinning in PCA Space

A further approach beyond light curve standardisations (such as SALT2) follows works on SNIa twinning [12]. This method assumes that spectrally similar SNIa (twins) have similar intrinsic brightness. These twins can then be used to compute distances and infer cosmology.

The Euclidean distance in PCA space,

$$d_{ij}^{PCA} = \sqrt{\sum_k^{15} (X_{i,k} - X_{j,k})^2} \tag{23}$$

where $\mathbf{X_i} = (X_{i,1} + X_{i,2} + ...X_{i,15})$ and $\mathbf{X_j} = (X_{j,1} + X_{j,2} + ...X_{j,15})$ are vectors of PCA coefficients for SNIa i and j respectively, is used to set these twins.

These twin pairs can then be evaluated through the magnitude residuals. For a pair of SNIa the residual in apparent magnitudes are given as

$$r_{ij} = \Delta m_{ij}^{obs} - \Delta m_{ij}^{model} \tag{24}$$

such that a value $r_{ij} = 0$ indicates perfect twins. Δm_{ij} indicates the difference in apparent magnitude observed (obs) and predicted ($model$) between SNIa i and j.

This methodology can then be extended to higher redshift SNIa since spectroscopic features can be recovered through PCA and compared, allowing for larger distance measures.

3.5 Twin Candidate Screening

Since this distance in PCA space is representative of spectral similarity, the closest pairs in PCA space are the most likely twin candidates when applying a full χ^2 twinning test [12]. Therefore, potential twins can be quickly identified in PCA space before full χ^2 analysis.

Since finding "twins" is rare this method is optimised to ensure a maximum or sufficient number of twins are found. This is achieved by testing the model over varying distance thresholds relative to a χ^2 percentile.

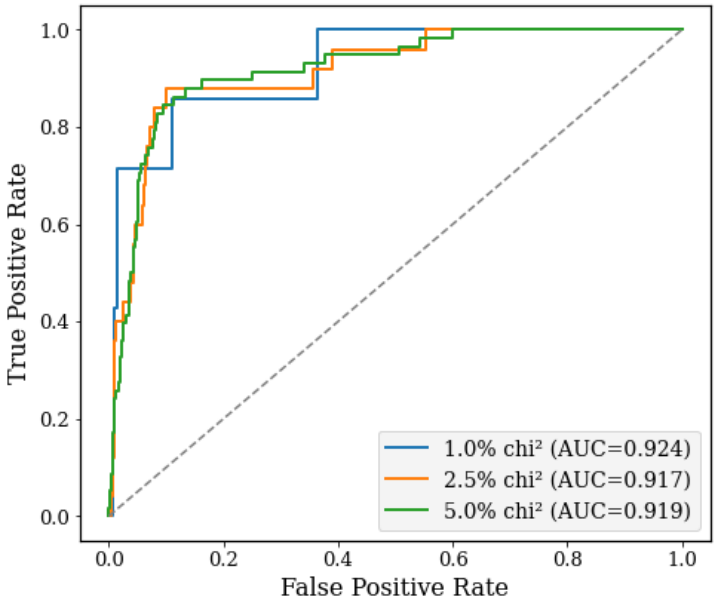

Figure 6: ROC curve indicating PCA distance screening for twins performance for identifying full χ^2 twins (set as truths).

The ROC curve 6 illustrates how the PCA distance thresholds are implemented. The true positive rate (TPR) must be high to ensure that sufficient twins remain. Since this is a screening procedure, the TPR should be prioritised while ensuring that limited false positives occur so that the process is worthwhile to reduce overall computational cost when implementing χ^2 twinning.

4 Results

4.1 PCA Representation of Supernova Spectra

This section demonstrates that PCA provides an accurate and low dimensional representation of SNIa spectra. Each dimension is represented through eigenspectra produced via PCA. The eigenspectra produced from PCA of the clipped dataset are illustrated in figure 7 below.

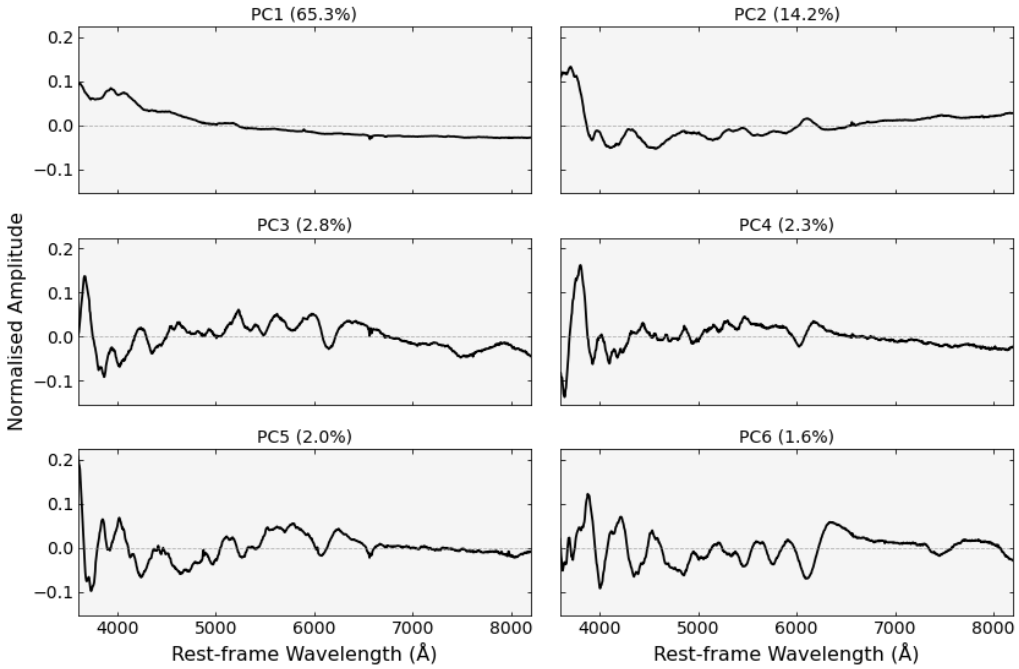

Figure 7: Illustration of the first 6 principal component eigenspectra from a clipped sample of low redshift SNIa spectra with relative variance encapsulated indicated.

As shown in figure 7 above, the first principal components (PC1 and PC2) are smooth with low variation over wavelength. These components capture overall global spectral properties, general continuum shape and broad relations such as color and temperature. Later PCs then contain sharper features, these indicate weaker absorption lines, rarer spectral features and spectral noise. Much of the spectral variation in the dataset is captured by the first few principal components, with 65.3% by the first alone, then 87% by the first six. PCs are ordered by explained variance rather than direct physical interpretability. Therefore, later PCs can still contain physically relevant information, but additional PCs show diminishing returns as shown by the 15 used illustrated in figure 8.

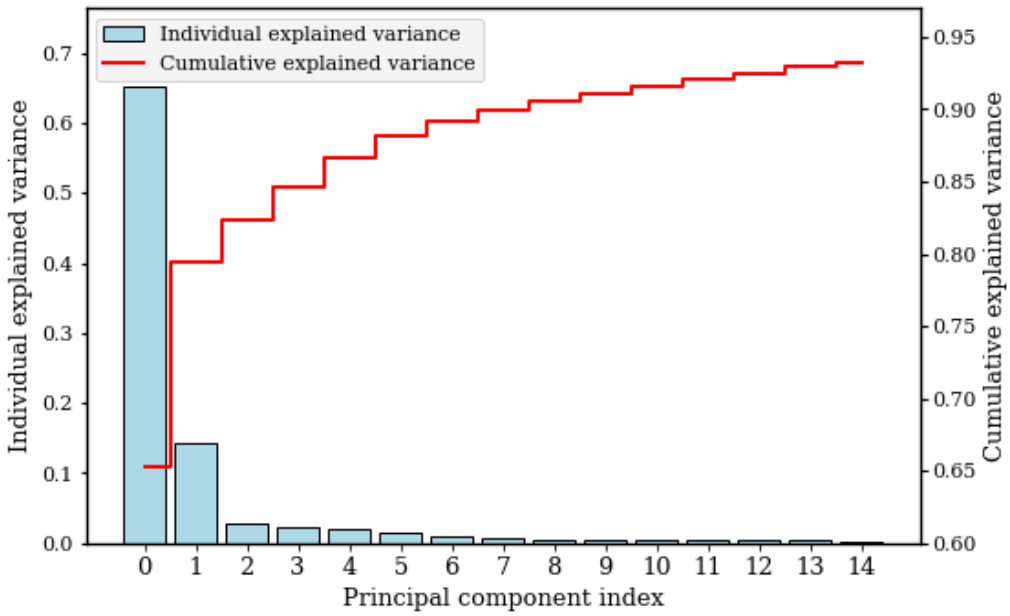

Figure 8: Explained dataset variance by each principal component.

SNIa spectra can then be reconstructed with various numbers of PCs used, illustrated in figure 9 below. As shown in the figure a small number of PCs, as low as 3, can reconstruct the overall shape of the SNIa spectrum allowing for major dimensional reconstruction and accounting for 77.5% of the datset variance. However, using this few PCs misses finer spectral details and can misrepresent absorption lines. Including further PCs up to 10 then distinguishes spectral lines more clearly and can more accurately depict spectra, representing almost 90% of the dataset's spectral diversity. Increasing again to 15PCs factored better represents the amplitude of spectral features and begins to show finer details, as shown at low wavelength in figure 9.

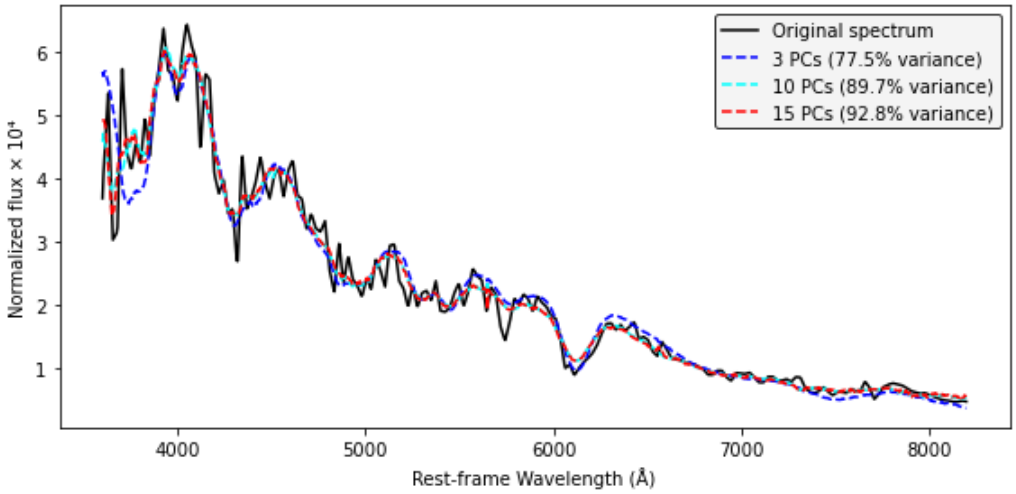

Figure 9: An example spectrum ZTF18aagstdc (normalised and shifted into rest-frame) overlaid with PCA decompositions using the first 3, 10 and 15 principal components.

Overall, figure 9 highlights that PCA can reconstruct spectra accurately with relatively few components and that the majority of spectral features are captured within 15PCs, providing potential for spectral comparisons and distance calculations.

4.2 SALT2 Extension Standardisation

PCA has potential to improve SNIa standardisation through a SALT2 extension, this improvement can be quantified through the Hubble scatter, the standard deviation σ_{Hubble},

$$\sigma_{Hubble} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \overline{r})^2}, \tag{25}$$

of Hubble residuals, r_i, the difference between the standardisations predicted distance modulus μ^{pred} and the distance modulus predicted by ΛCDM cosmology, μ^{cosmo}

$$r_i = \mu_i^{pred} - \mu_i^{cosmo}. \tag{26}$$

Figure 10 highlights how additional PCA coefficient standardisation terms can improve SALT2 based standardisation.

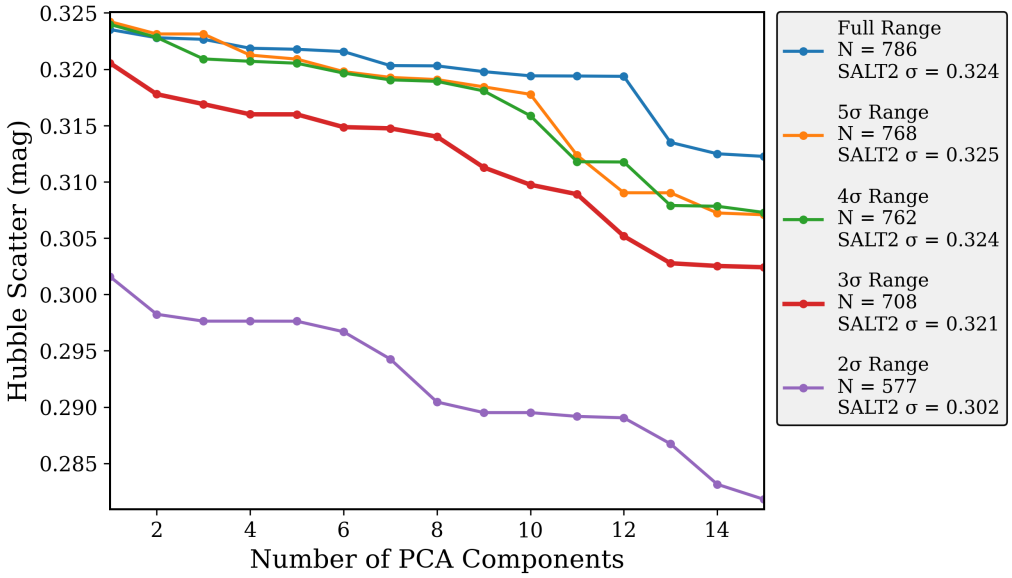

Figure 10: Hubble scatter of Type Ia supernovae as a function of the number of PCA components used to correct SALT2 distance moduli. Lines correspond to different standard deviation clipping levels applied to the PCA coefficients, SNIa with any coefficient outside of these ranges are removed. The scatter is the standard deviation of residuals relative to the theoretical distance modulus from a Flat ΛCDM cosmology.

As shown in figure 10 as the number of PCs factored increases, scatter reduces for all PC clipping ranges. Additionally, higher PCs also contribute to scatter reduction despite representing low variance in the dataset, but scatter reduction encounters diminishing returns for higher PCs 13-15 and beyond. Applying clippings to the dataset reduces the spectral diversity and reduces scatter, however, this is balanced by dataset reduction. The 2σ clipping range offers a large scatter reduction of $0.021\,mag$ compared to the 3σ range but reduces the number of SNIa significantly. The 3σ range retains 90% of the original sample whereas, the 2σ range only retains 73%.

A 3σ clipping range is used for further analysis, providing a reduction in scatter from $0.321\,mag$ scatter for SALT2 only to a $0.302\,mag$ as shown in figure 11 below.

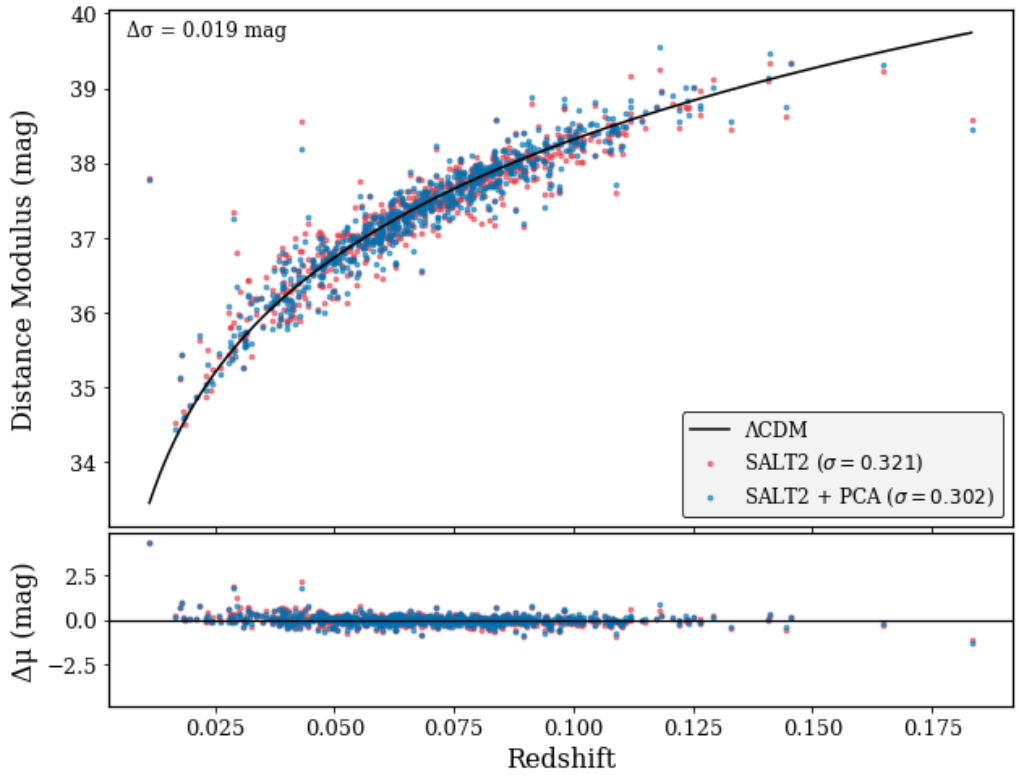

Figure 11: Hubble diagram with residuals panel (bottom) indicating the improvement in distance modulus scatter when extending SALT2 SNIa standardisation with PC coefficients. This plot is constructed from a dataset clipped to a 3σ standard deviation range of PC coefficients. σ in the plot refers to the standard deviation of residuals. The PCA extension factors the first 15 PCs and is compared to the fitted ΛCDM cosmology.

The Hubble diagram comparing SALT2 with PCA extension shown in figure 11 measures distance modulus across redshift. Introducing this PCA extension shows a reduced dispersion around the fitted ΛCDM model. The residual plot (bottom panel) shows that for the extension there is a tighter fit around the zero point across the entire redshift range indicating that this extension improves scatter across varying redshift, reducing the average scatter by $0.019\,mag$ compared to SALT2 alone.

Overall, it is shown that extending SALT2 standardisation using PCA coefficients measurably and consistently improves SALT2 standardisation. Crucially, the scale of this improvement is dependent on the number of PCs and dataset clipping but additional terms consistently reduce Hubble scatter.

4.3 Candidate Twin Identification

In order for twin screening to be quantified, a spectral squared distance is defined as

$$SSD_{ij} = \sum_{\lambda} \left(f_i(\lambda) - f_j(\lambda)\right)^2 \tag{27}$$

where λ represents interpolated wavelength values and f the associated flux at those values for compared SNIa i and j. This metric is analogous to a χ^2 metric without errors factored,

and therefore the typical twinning [12]. As such twinning in this space is used as the truth throughout this analysis, and PCA screening a predicted twinning measure.

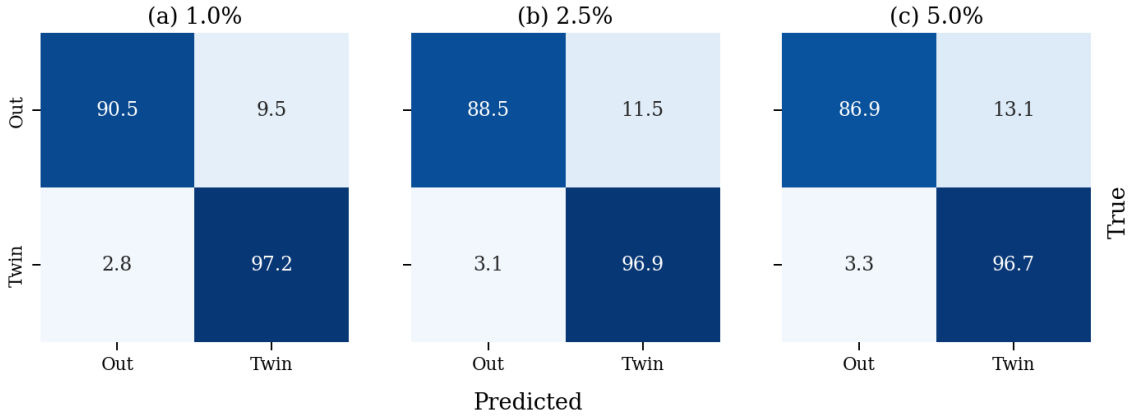

Figure 12: Confusion matrices showing the performance of PCA-based twin screening for different squared spectral distance percentiles (1%, 2.5%, 5%). True twins are pairs within the lowest percentile of squared spectral distance, defined by equation 27. Rows correspond to the true classification, while columns indicate the predicted classification based on PCA distance. Values are shown as row-normalised percentages using a Youden-optimised threshold, balancing twin recovery and limiting false positives.

The predictive performance of PCA-based twin classification is shown in figure 12, where PCA distance thresholds are Youden optimised to balance the high recovery rate of true positive twins with minimal false positive twins identified. True, spectral squared distances of 1%, 2.5% and 5% are shown, where true twins are SNIa pairs with spectral squared distances within the lowest percentage threshold indicated. Across all comparisons shown PCA-based screening recovers 97% of true twins with a small 10-13% of non twins falsely identified. Therefore, this metric shows an efficient screening approach without major accuracy loss. However, since twins are relatively rare this approach should also be optimised to identify a maximum number of true positive twins. The results for this optimisation are shown in figure 13.

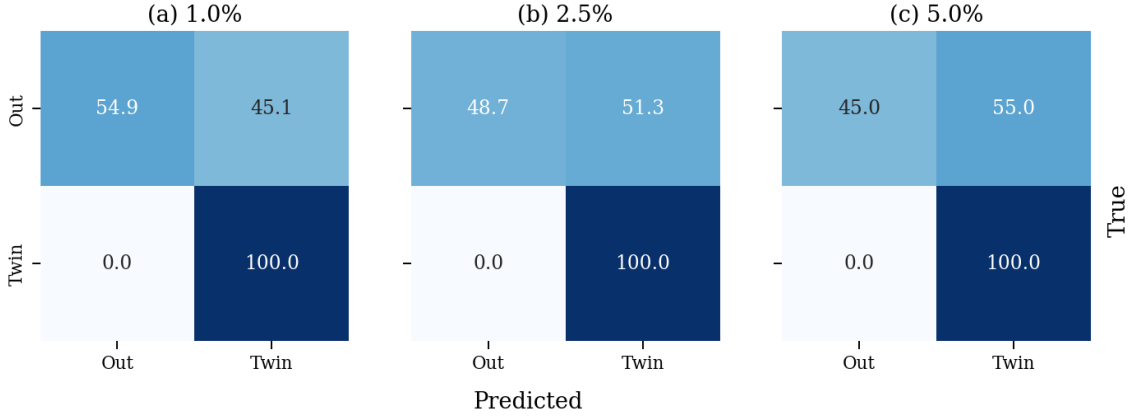

Figure 13: Confusion matrices showing the performance of PCA-based twin screening for different squared spectral distance percentiles (1%, 2.5%, 5%). True twins are pairs within the lowest percentile of squared spectral distance, defined by equation 27. Rows correspond to the true classification, while columns indicate the predicted classification based on PCA distance. Values are shown as row-normalised percentages using a threshold optimising the true positive rate to recover the maximum number of twins.

When optimising for the maximum number of true twins recovered, 100% can be recovered across all 1.0, 2.5 and 5.0% thresholds. However, this method also finds between 45.1% (panel (a) figure 13) and 55.0% (panel (c) figure 13) false positives at 1.0% and 5% thresholds respectively. This highlights that PCA distance based screening can successfully identify all twins across varying thresholds, however, this is less efficient as an additional 50% are typically falsely identified as twins.

4.4 PCA Distance Distributions

The PCA distance, d_{ij}^{PCA} (equation 23) is a Euclidean distance while the twinning spectral distance is a pseudo-χ^2 spectral similarity measure. To assess similarity in the two distance distributions, the spectral squared distance, SSD,(equation 27) must be adjusted to,

$$d_{ij}^{spec} = \sqrt{SSD_{ij}}, \tag{28}$$

the same geometrical form as the PCA distance, facilitating the comparisons shown in figure 14.

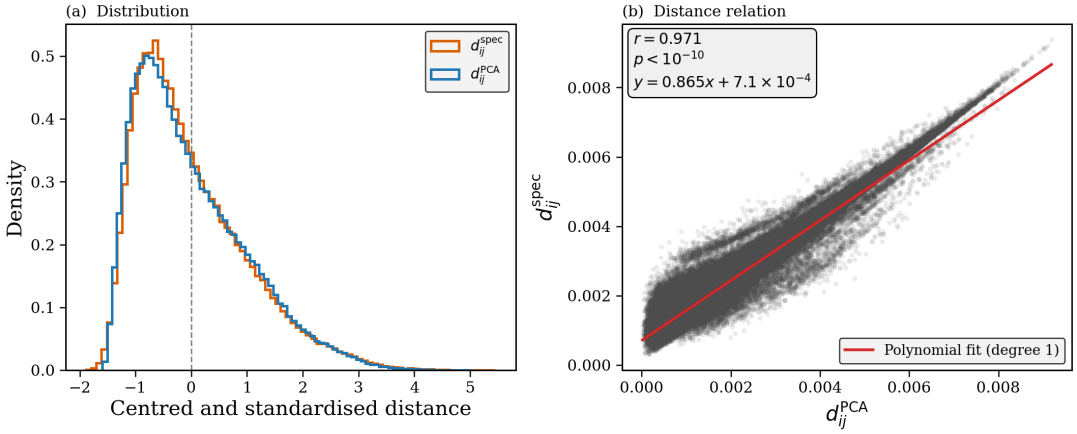

Figure 14: Comparisons of the spectral similarity distance metrics d_{ij}^{spec} (equation 28) and PCA distance d_{ij}^{spec} (equation 23). Panel (a) shows the standardised distribution of the distances centred at zero, showing close agreement and indicating that PCA preserves the global structure of pairwise spectral distances. Panel (b) shows the relationship between the two distance measures, demonstrating a strong correlation (r=0.97) and an approximately linear fit, with PCA distances providing a scaled representation of spectral distances.

Panel (a) demonstrates that the centred and standardised distributions for both d_{ij}^{PCA} and d_{ij}^{spec} metrics show close agreement. Both have similar shape, peaking at negative standardised values and have long positive tails. The SNIa are therefore very similarly distributed across both spaces.

Panel (b) shows a strong positive correlation (r=0.971) between the two metrics. This linear fit extends across the full range of spectral similarity. The two metrics are related through a best fit polynomial

$$d_{ij}^{spec} = 0.865\, d_{ij}^{PCA} + 7.1 \times 10^{-4}. \tag{29}$$

The majority of SNIa poinys lie at low distance in both metrics, in a densely clustered region. This is expected indicating that SNIa are broadly similar.

Ultimately, there is a strong correspondence between PCA distance and the converted pseudo-χ^2 twinning measure, making PCA distance a valid spectral similarity measure. This further justifies the use of PCA distance in SNIa twin screening techniques or as a basis for a twin standardisation.

4.5 Twin Based Standardisation

This method assumes spectral similarity between SNIa pairs indicates similar spectral luminosity, this spectral similarity is quantified in PCA space. To evaluate this method scatter residuals are compared, since the SNIa are compared in pairs the residual, r_{ij} becomes

$$r_{ij} = \Delta m_{ij}^{obs} - \Delta m_{ij}^{model} \tag{30}$$

where Δm_{ij}^{obs} is the observed magnitude difference and Δm_{ij}^{model} is the magnitude difference predicted by the ΛCDM model for SNIa i and j. To determine this scatter across a sample of N pairs the RMS residuals are determined as

$$RMS = \sqrt{\frac{1}{N}\sum_{i,j} r_{ij}^2} \tag{31}$$

hence a lower RMS value indicates a better standardisation. Figure 15 indicates this RMS scatter across varying percentage thresholds of "closest" SNIa in PCA space.

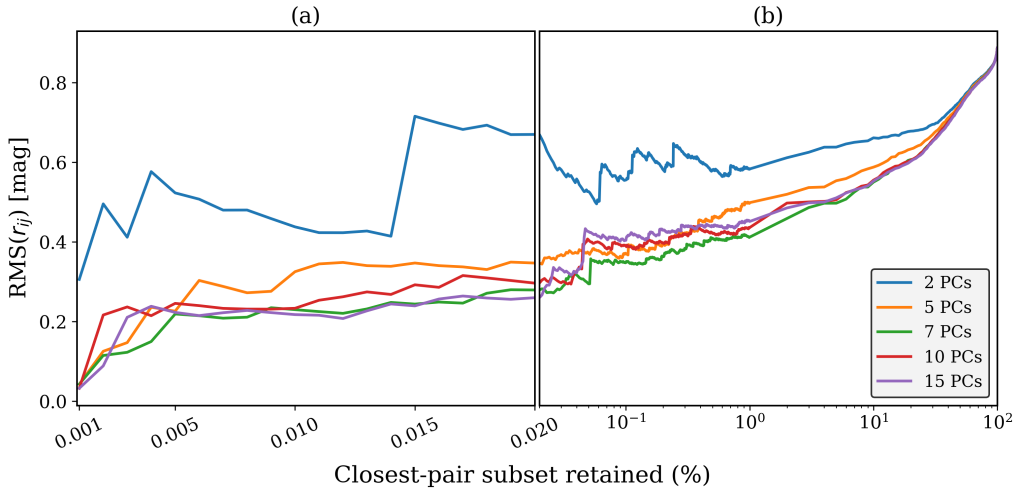

Figure 15: An indication of the RMS magnitude scatter values (equation 31) evaluated over varying percentile thresholds of lowest PCA distance determined from PCA with 2,5,7,10,15 components. Panel (a) shows the low percentile range of the closest 0.02% of SNIa pairings with a linear scale. Panel (b) illustrates the remaining 0.02% to 100% range with a $\log_{10}$ percentile scale.

As shown in figure 15 RMS scatter progressively decreases as closer SNIa are compared. There is then a significant reduction in scatter at low percentiles, illustrated in Panel (a) showing SNIa with extreme spectral similarity or "Twins". The initial comparison of all possible pairwise SNIa comparisons yields an RMS scatter of $0.887\,mag$. Reducing the factored pairings to the closet 1% of pairings RMS scatter significantly reduces to $0.412\,mag$ and $0.3469\,mag$ at 0.1%. At an extreme selection of 0.01% the scatter is $0.217\,mag$ however, this only factors 22 SNIa so a larger dataset would be required to verify this scatter reduction. Further reductions to the

closest 2 and 4 twins achieves an RMS scatter of 0.0324 mag and 0.0343 mag though due to the small sample size, this is not statistically robust. Across all pairings and the closest 0.010% of pairings using the first 7 PCs achieves the lowest scatter reduction so this number is used for further analysis.

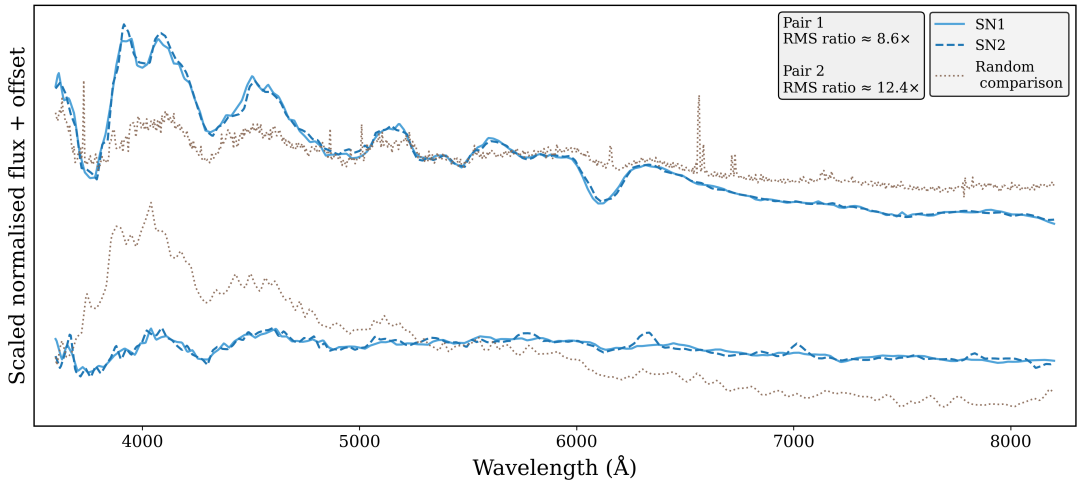

Figure 16: Spectral comparison of close proximity PCA space SNIa pairs with random SNIa using normalised, interpolated spectra and 7PCs. Pair 1 (top) shows a 8.6 times larger RMS between the random spectra (ZTF18abdhsfc) and SN1 (ZTF20aautscv) than SN1 and SN2 (ZTF20aauuiri). Pair 2 (ZTF18aanygjk & ZTF19acmbjvi) has a 12.4 times lower RMS than ZTF18aanygjk and a random SNIa (ZTF20abjnmhn).

Figure 16 shows examples of low PCA distance, twinned SNIa compared to a random SNIa. In both examples the twins overlap closely over the full wavelength range while the random SNIa show clear deviations in spectral feature shape and magnitude. Quantitatively comparing the twin pairs shows RMS values 8.6 and 12.4 times lower relative to the random SNIa shown.

Table 2: Quantitative results from spectral comparisons shown in figure 16, where "twin" subscript refers to SN1-SN2 comparisons and "random" refers to SN1-random comparisons. Correlarion ($Corr$) is the calculated pearson correlation between SNIa spectra and PCA Distance is determined using 7PCs.

Pair	RMS_{twin}	RMS_{rand}	$Corr_{twin}$	$Corr_{rand}$	PCA Distance	Twin Residual (mag)
1	1.1×10^{-5}	9.7×10^{-5}	0.997211	0.898673	0.000111	0.0510
2	1.2×10^{-5}	14.7×10^{-5}	0.924514	0.467042	0.000117	0.0322

The absolute RMS values shown in table 2 of order 10^{-5} for twins show extreme spectral similarity, an order of magnitude lower than the RMS for the random comparison. The correlation values, particularly for twin pair 1, of 0.997 indicates almost identical spectral behaviour while the correlation with random SNIa quantifies clear variation. The corresponding scatter residuals between twins is also very small with sub 0.06 mag values.

Table 3: Redshift values for low PCA distance "Twin" SNIa pairs from figure 16.

SNIa ztfname	Redshift
ZTF20aautscv	0.05931
ZTF20aauuiri	0.05680
ZTF18aanygjk	0.06181
ZTF19acmbjvi	0.09238

Redshift values of 0.059 and 0.057 for SNIa pair 1, shown in table 3 are within a narrow redshift range. Pair 2 has a larger, significant redshift difference of $\Delta z = 0.03$, indicating that spectrally similar SNIa appear with both similar and distinctly differing redshift values.
In general twinning in PCA space can identify spectrally similar SNIa across varying redshift and lower PCA distances are associated to lower residuals.

5 Discussion

5.1 Effectiveness of PCA for Spectral Representation

In order to efficiently standardise using SNIa spectra, they must be decomposed to a smaller number of parameters. PCA has been shown to compress SNIa spectra with sufficient variation captured and also capture underlying physics. PCs 1 and 2 show smooth, continuous structure as shown in figure 7, capturing global properties such as temperature and color. Later PCs contain sharp lines and prominent features, encoding absorption lines and rare events.

Much of the SNIa variation is captured within the first few PCs with PC1 capturing 65% of the overall variation of the dataset, achieving 90% with the first 10. Therefore, much of the SNIa spectral information lies in a low dimensional manifold in PCA space enabling significant decomposition from full spectra, PCA-based standardisations (3.4.1) and matching SNIa spectra (3.4.2). Although higher order PCs encapsulate less variance, they can still encode physical features, particularly rare spectral events, however, these later PCs are more susceptible to noise. Therefore, since spectral variation is mostly captured within the first few PCs, most of the spectral lines are present when including 10-15PCs and additional PCs provide diminishing returns and increase noise, it is valid to truncate the number of PCs to varying levels below 15 depending on application.

It is important to note that this approach uses a linear model sensitive to preprocessing. The model can therefore be refined to factor physically meaningful wavelength regions such as the Si II 6355$\mathring{A}$ absorption feature and Fe-group dominated regions [33], [34] and more advanced preprocessing techniques such as Gaussian processed applied. In general, the methodology accurately describes SNIa spectra through a small number of coefficients applicable for standardisation.

5.2 PCA-Based Spectral Similarity and Twin Identification

A defined PCA distance (equation 23) is evaluated as a spectral similarity measure through comparisons with a pseudo-χ^2 spectral similarity measure (equation 27).

Assessing the PCA-based SNIa twin candidate screening procedure, approximately 97% of true twins are correctly identified across all 1.0, 2.5, 5.0% true twin thresholds, with a Youden optimised PCA distance threshold. This process achieves a low rate of 10-13% of falsely identified twins. This demonstrates that PCA distance retains the majority of physically similar SNIa pairs while substantially reducing the number of candidate pairs.

Adjusting PCA distance thresholds also provides dynamic balance between completeness and efficiency of pre-screening samples. For twin discovery, a maximum number of true twins should be identified, altering the threshold allows 100% recovery of true twins. This full

recovery is balanced with significant sample contamination with approximately 50% of non-twins incorrectly identified as twins. Therefore, for efficient twins analysis, the Youden optimised threshold minimises contamination, providing a computationally efficient SNIa twin screening technique.

Further comparisons of the two spectral similarity measures show strong geometric similarity. PCA distance and the pseudo-χ^2 metric have a strong correlation (r=0.97) and near linear relationship across the full range of distances. PCA distance is therefore shown to be a scaled spectral similarity measure. Additionally, PCA distance preserves the global structure, ordering of spectral similarity and relative spectral similarity between SNIa.

Ultimately, the screening success shown highlights that the PCA and pseudo-χ^2 distance measures show strong correspondence, with nearest neighbours in PCA and flux spaces largely consistent. Since PCA compresses spectra into dominant modes, such as continuum shape and major absorption features, which dominate spectral similarity, PCA not only provides a statistically sound similarity metric, but also encodes dominant physical spectral features. Specifically, the majority of SNIa are observed in low distance clusters for both distance metrics, consistent with the known spectral homogeneity of SNIa [35]. However, PCA distance still allows spectral comparisons within this spectrally similar cluster, facilitating SNIa twin-based standardisation.

The method does have several limitations, first, the pseudo-χ^2 metric introduced to identify "true" twins is unweighted, factoring no errors and is therefore not a full χ^2 similarity measure. This PCA distance is similarly unweighted, and sensitive to truncation of PCs used and pre-processing. The twin definitions in this work are simply percentile based and so somewhat arbitrary labels, a physics based threshold in PCA space would be required for a physically motivated twin classification.

Overall, these results support PCA distance as a robust, computationally efficient proxy for spectral similarity, which is readily applicable to large datasets. Supported by the high recovery of twin identification with controlled contamination, PCA distance provides both an effective pre-selection tool for full χ^2-based spectral twinning and a foundation for a twinning-based standardisation.

5.3 Standardisation Techniques

5.3.1 PCA-Extended SALT2 Standardisation

It has been shown that extending SALT2 standardisation with PC coefficients reduces scatter relative to SALT2 standardisation alone. An extension using 15PCs on a dataset within 3σ ranges of all PC coefficients can provide a consistent improvement of $0.02\,mag$. While this is a modest scatter improvement of 6%, the method does show potential as an extension to this standardisation approach. Fundamentally, SALT2 is a low dimensional standardisation approach which cannot capture full SNIa diversity. Introducing a PCA extension from SNIa spectra, these higher dimensions capture variance in absorption features, line strengths and subtle spectral features. Ultimately, PCA can therefore account for variation in SNIa not yet factored by SALT2 approaches.

Each PC added factors more spectral variation but with diminishing returns. Higher PCs encapsulate less spectral diversity and can account for noise. Analysis has shown that scatter reduction typically plateaus at 13-15PCs or beyond. Early PCs capture global spectral shape and are the most important features to account for as they capture the most spectral diversity. Later PCs typically account for noise and rarer features. Therefore, limiting this extension to 15PCs ensures that sufficient spectral variation is factored while preventing overfitting and

allowing noise to impact standardisation.

The spectral diversity of the SNIa is also shown to affect standardisation. Applying clippings to the dataset in PCA space such that PC coefficients must lie within deviation levels of the average outliers are removed. A 3σ range is used as the standardisation improves at this level while retaining over 90% of the original dataset. Reducing the range to 2σ improves standardisation but loses 25% of the original sample. Since this method relies on a balance of improved standardisation and sufficient sample size the 3σ deviation range is justified. Tighter clippings do result in a lower scatter since the dataset becomes more homogeneous, but crucially PCA-based extension improves standardisation across all clipping ranges. Reducing this range to 1σ (shown in Appendix B.1, figure 17) results in a larger 17% scatter reduction from SALT2 standardisation alone on this clipped dataset of $0.192\,mag$, highlighting how PCA space can also be used to identify spectrally similar SNIa. After applying a 15PC extension to the standardisation the scatter drops to $0.159\,mag$. However, the sample size significantly drops to 115 SNIa, an 85% reduction. This does show the potential for a further standardization method clustering the sample into subgroups in PCA space before applying the PCA standardisation extension with aims to achieve this higher scatter reduction across the entire dataset.

In the 3σ clipping range used, the method achieves a 6% scatter reduction of $0.02\,mag$, highlighting the effectiveness of this PCA methodology for SNIa standardisation, however, the scatter reduction is modest. The method does provide several advantages, first it utilizes a large proportion of the survey dataset, it is scalable and coefficients are physically connected through the spectra and further spectral analysis can provide higher physical insight. The method is also modular and additive since this PCA extension can be applied after other SALT2 extensions (such as host mass corrections [36] or morphology [37]) and doesn't replace any prior variables including the SALT2 stretch and color terms. This method should therefore be seen as a higher order correction layer aimed to capture the remaining variance.

There are, however, several limitations that must be noted, currently this analysis assumes a linear model, depends on preprocessing and requires relatively high signal to noise in the SNIa spectra. The method does provide multiple avenues for further investigation such as factoring the full spectral time series rather than peak luminosity spectra, applying a non-linear PCA model and extending to higher redshift.

5.3.2 PCA-Based Spectral Twinning

PCA distance provides a quantitative metric and identifier of spectral similarity therefore physical similarity between SNIa. Since the greater physical similarity two SNIa have, the closer their intrinsic brightnesses are and lower their Hubble residuals, finding these extremely similar SNIa (twins) provides the potential for better standardisation. SNIa spectra encode ejecta conditions including temperature, density, velocity and composition, through spectral lines such as Si II λ6355, Ca II H&K, the Ca II near-infrared triplet, and Fe II/Fe III blends, all of which trace the thermodynamic and ionisation state of the ejecta and therefore influence the intrinsic brightness of SNIa.

PCA distance acts as a compact description of spectral variation and twinning in this space allows identification of globally similar spectra. Compared to typical light curve standardisations, spectral analysis can capture physical properties more directly. Using PCA distance as a twins identifier, the method can be validated through RMS between spectra,

an identifier of absolute spectral difference, and Pearson correlation values, an indication of spectral shape similarity. Overall, lower PCA distances for pairwise SNIa comparisons result in lower RMS values and higher correlation. The closest two pairs, in 7-dimensional PCA space, for the dataset, have RMS values of 1.1×10^{-5} and 1.2×10^{-5}, an order of magnitude lower than comparing these SNIa to another at random. The correlations also show excellent agreement with pair 1 having a Pearson correlation values of 0.997 indicating almost identical spectral shape. These twin pairs also produce low residuals of 0.0510 and $0.0322\,mag$ substantially lower than the average pairwise comparison residual of $0.887\,mag$. However, due to the small number of SNIa pairs at this 0.002% closest percentile, further analysis is required to test if the observed $0.03\,mag$ residual values can be achieved consistently. Identification of the top 1% closest pairs does show a significant improvement with RMS residual values of $0.412\,mag$ across a sample of 2268 SNIa, a greater than 50% scatter reduction.

Twinning in PCA space using 7 PCs produces the lowest average scatter across the entire percentile range and the lowest average scatter in the closest 0.010% of twins, as illustrated in figure 15. Performing a 5-fold cross validation test to test how many PCs used for twinning reduces the scatter most significantly (figure 19, Appendix C) again shows that 7PCs achieve the lowest scatter across the dataset, highlighting that leading components capture much of the spectral variation. At lower percentiles, such as 5%, factoring a larger number of PCs does reduce scatter, however, they likely begin to factor noise, reducing the robustness across all pairwise comparisons.

Looking back at the closest pairs 1 and 2 in 7PC space, the redshift differences of the pairs is $\Delta z = 0.00251$ and $\Delta z = 0.0306$ respectively. This significant difference in redshift range indicates that twins can be identified at both similar and distinct redshifts, therefore the method captures intrinsic SNIa properties despite observational effects. This provides a promising extension looking at twinning high redshift SNIa to low redshift pairs to infer cosmology since analysis of high redshift SNIa is required to differentiate cosmological models. This method provides several avenues for future work in additional to high redshift applications including full time series analysis of spectra, defining a twins metric in PCA space to define SNIa twins and applying standardisations to twinned SNIa such as SALT2 color parametrisations.

The method does have several limitations at this stage. First, the PCA assumes linearity similarly to all methods discussed, though this can be tested through varying PCA procedures and only one spectra is evaluated not the full time series. The method also relies on spectra with relatively high signal to noise, which are increasingly difficult to produce in higher redshift regimes.

5.3.3 Comparison of Standardisation Methods

Both standardisation methods outlined (SALT2 PCA extension 3.4.1 and spectral twinning in PCA space 3.4.2) aim to improve SNIa standardisation by factoring spectral information through PCA decomposition. The SALT2 extension method provides a global parametric correction across the entire dataset while the twinning method involves pairwise matching. These differing standardisation approaches prevent straightforward quantitative comparisons though the differences and tradeoffs can still be discussed. The methods define residuals differently therefore comparing scatter results is not valid.

Starting with the SALT2 PCA extension, this method extends the typical SALT2 correction to

include spectral information, redefining distance modulus μ as defined in equation 22, providing a global fit across the entire SNIa sample. The method is extendable and can be applied readily to Hubble diagrams to infer cosmology as shown in figure 11. The PCA twinning method matches SNIa by distance in PCA space, operating on the assumption that similar spectra correlates with similar intrinsic luminosity. With this approach physical similarity of SNIa is directly evaluated in pairwise comparisons. The key differences between the two approaches is that the SALT2 based standardisation is a global parametric fit, assuming a linear model, while twinning is a nearest neighbour standardisation requiring no explicit model.

Ultimately, both methods improve standardisation, with the SALT2 PCA extension reducing the overall Hubble scatter across the full dataset achieving a 6% scatter reduction across a large statistically stable dataset. Twinning meanwhile achieves high precision locally with potential for scatter between pairs at sub $0.05\,mag$ levels. Since the SALT2 PCA extension provides a global, scalable, regression based standardisation ready for cosmology and twinning acts as a physically motivated, local refinement with high redshift potential, the two methods should be viewed as complimentary approaches for SNIa standardisation. Both extract useful spectral information through PCA and have shown standardisation improvements despite fundamentally different methodologies and underlying assumptions.

5.4 Implications for Cosmological Distance Measures

Fundamentally SNIa standardisation is conducted to determine more accurate distance measures providing more accurate cosmological insight. From these distance measures, cosmological parameters such as the Hubble constant, H_0, and the matter density parameter, Ω_m, are constrained. Reduced residuals (as achieved by both outlined methods) enables these parameters to be constrained more accurately and cosmological models to be tested. Supernova cosmology requires a balance of precision measurements and sample size to ensure meaningful and statistically sound results are produced. The PCA twinning method outlined offers a high-precision, small sample size approach while the SALT2 extension provides an improved standardisation across the entire dataset. Therefore, together these approaches house potential for improved cosmological inference. However, while improved standardisation aims to provide more accurate inferences, improved standardisation doesn't guarantee tighter constraints (see Appendix B.3) since for example in the PCA extension, up to 15 additional parameters are introduced all carrying uncertainty. While these additional terms can introduce further uncertainty in cosmological parameters such as Ω_m, this is a reasonable trade-off given the improvement in standardisation accuracy and physics factored in standardisation.

5.5 Robustness & Limitations

5.5.1 Reconstruction Robustness

PCA spectral reconstruction is assessed by limiting the seen wavelength, decomposing the full spectrum through PCA and comparing the true hidden region with that extrapolated through PCA (see appendix D.1). The PCA method outlined successfully reconstructs these hidden regions. Since SNIa are known to exhibit correlated structure throughout the spectrum with their diversity well-described by a low-dimensional representation [33], [14], the ability to reconstruct unseen regions demonstrates that this PCA-based spectral deconstruction is not merely simplifying spectral patterns but capturing physically driven spectral relations in reduced dimensions. These finding support the use of low dimensional spectral interpretations in SNIa standardisation.

While showing successful reconstruction in general, reconstruction quality is dependent on the

seen wavelength range. Blue wavelength regions ($3600 - 4600\mathring{A}$) provide the best performance. This is expected as this region encodes large spectral variation and key features so contains high spectral information and requires more to be successfully reconstructed. Visually, key features are reconstructed effectively, element-dependent spectral lines are predicted, highlighting that the methodology successfully captures physical diversity.

Quantitative analysis of this reconstruction performance introduces two metrics, Pearson r, encoding spectral shape recovery, where r=1 indicates identical spectra. The derivative of SSD (equation 28), quantifies the success of feature and line reconstruction, with lower values indicating better feature recovery. Across a random sample of 15 spectra, the typical Pearson r value of hidden regions is 0.984 reaching 0.997 indicating almost perfect recovery can be achieved. Even the lowest value of r=0.731 indicates reasonable shape recovery. Therefore, despite significant dimensional reduction to 15PCs, the overall global description of SNIa spectra was effectively captured, indicating that the primary spectral diversity of SNIa is governed by a small number of underlying physical parameters encoded within the leading principal components. Evaluating the recovery of spectral features through SSD derivative values ($d(SSD)$), the previous highest performing SNIa has a low $d(SSD) = 8.5 \times 10^{-14}$ with typical and poor values of 14.9×10^{-14} and 305×10^{-14}, indicating that while spectral structure is strongly encapsulated, the extrapolation of fine details varies. Observationally, "typical" SNIa perform much better, these SNIa have finer details well captured by the PCs hence are reproduced better. This is therefore another motivation for clustered spectra in PCA space allowing deeper physical representation of less "typical" spectra.

The results show that spectral features are not independent across wavelength with key information encoded across the spectrum. PCA captures this information and physically meaningful variation. Since the majority of spectral reconstructions are high quality but some are significantly worse, these poorer reconstructions likely indicate the original spectra are noisy or contain rare spectral behaviour (shown in visualizations e.g. figure 21). These poor cases feature misaligned features and incorrect line depths indicating limits to this linear PCA model.

This PCA distance-based methodology is robust to incomplete spectra validating its use for high redshift SNIa and heterogeneous datasets. Therefore, PCA-based standardisations and PCA space spectral twinning methods are justified, and this robustness measure supports the underlying assumption that similar PC coefficients represent similar spectral similarity. The method is dependent on the wavelength range and spectral quality, and PCA can fail for outlier SNIa spectra.

Overall, PCA captures both global and feature level spectral structure, with blue wavelengths dominating the information content. The method is robust, but not universally reliable.

5.5.2 Twinning Robustness

To evaluate the PCA-based twinning techniques, artificial noise is injected to SNIa spectra and twin pairs, ordering distributions and retained SNIa in percentiles compared to that for the original spectra (see Appendix D.2) for full noise explanation). As the noise models applied are illustrative rather than physically calibrated to a specific survey, the analysis is interpreted qualitatively, focusing on the relative stability of PCA-based twinning.

PCA distanced-based twin identification remains stable under noise, with closest percentile SNIa largely unchanged. Twin selection is therefore driven by intrinsic spectral structure.

Varying noise models reflect different physical mechanisms in spectroscopy. Gaussian and correlated noise approximate typical detector and calibration effects. Perturbations such as spectral tilts and localised distortions mimic flux calibration errors and instrumental artefacts.

Poisson like noise replicates the impact of low signal to noise and low photon count SNIa observations.

Twinning remains largely consistent with the injection of Guassian and correlated noise models. This validates this twinning approach as these noise levels are typical spectroscopic affects which frequently appear. Under chunk noise twinning is relatively stable. However, this model is largely dependent on which spectral windows are distorted, with distortion of higher information ranges causing more significant affects. Tilt noise causes moderate impact on twinning, and using more PCs typically results in lower impact with this model. Finally, Poisson based noise introduces catastrophic impact on twinning, highlighting the major signal to noise limitation common across spectroscopic techniques.

While the method remains largely robust to moderate noise levels, motivating the use in twinning high redshift SNIa with a low redshift training set, reasonable signal to noise is essential for the methodology. As signal to noise drops with increasing redshift, twinning certain SNIa becomes unviable when limited by spectroscopic resolution at extreme redshift or contamination introducing environments.

5.5.3 Limitations

Although the standardisation methods outlined show a reduction in residuals, both have several limitations. These limitations arise primarily in three forms; data, method and physical insight.

First, the dataset itself imposes limitations, the sample size faces a number of cuts and restrictions, such as wavelength limits and spectra used within 2 days of peak luminosity, therefore the sample used may not fully represent overall SNIa diversity. Additionally, only high signal to noise spectra can be used, this favours lower redshift SNIa in low contamination environments. Applied cuts additionally limit redshift and low redshift SNIa dominate the sample. The dataset therefore forces the assumption of relative SNIa homogeneity across redshift and environment.

The methods also impose limitations. Both require PCA to be conducted and spectra to be preprocessed. This preprocessing, particularly the interpolation of spectra, may result in artificial correlations or the loss of physical features. PCA is linear and variance driven, therefore nonlinear relations and rare but important features can be lost. The number of PCs chosen also has significant impact on the results, too few results in insufficient information captured and too many begin to introduce noise, so choices must be statistically robust. Method specific limitations also appear, the SALT2 PCA extension assumes a linear relationship between luminosity and PCs. Twinning is based on Euclidean distance, assuming all PCs are equally meaningful. Also, not all SNIa have twins and some have multiple therefore this method discriminates against unusual SNIa without twins. Finally, twinning relies on the assumption that similar intrinsic luminosity results in lower distance in PCA space, this assumption has not been explicitly tested using explosion simulations and relies instead on the assumption that SNIa with similar spectra have similar intrinsic luminosities.

5.6 Improvements & Future Work

To improve the outlined standardisation methods, the full spectral time series should be factored rather that only one peak luminosity spectrum per SNIa. Non-linear dimensional reduction such as autoencoders [38] or kernal PCA [39] should also be applied. Assumptions should be tested more rigorously, including factoring explosion physics and further analysis should be done in order to determine the optimal number of PCs balancing noise with valuable information obtained from spectra. The preprocessing pipeline can also be refined, particularly

spectral interpolation, which could use more complex methods such as Guassian processes [40]. These methods show potential for future work.

Spectral twinning through PCA analysis naturally extends to find **high redshift twins** (see Appendix C.1) as spectral decomposition trained on low redshift SNIa projected onto high redshift SNIa naturally filters out noise introduced through extrinsic factors. This allows spectral twins over larger redshift to be found and intrinsic luminosities of high redshift SNIa to be inferred. This approach is particularly valuable for cosmological inference as higher redshift SNIa are required to differentiate cosmological models.

Rather than undertaking one PCA, multiple **nested PCAs** could provide improved representation of SNIa and identify sub populations. A pipeline can filter and separate SNIa into different groups then conduct PCA again in these groups to evaluate further spectral diversity.

Since the SALT2 PCA extension shows significant scatter reduction in low PC coefficient variation subsets (e.g. $\leq 1\sigma$), clustering by PC coefficients can create small sub populations of spectrally similar SNIa. Then **fitting standardisation coefficients on clustered SNIa** has potential to achieve this significant scatter reduction across the entire dataset.

Twinning can also be refined by **weighting PCA distance**, particularly if done to reflect the physical significance of each PC with regards to intrinsic luminosity. This would then remove the assumption that each PC affects intrinsic SNIa luminosity equally.

Finally, the PCA SALT2 extension and PCA twinning approaches should be combined to create a method that can both isolate low intrinsic scatter between very similar SNIa and provide a scatter reduction across a large dataset of SNIa.

6 Conclusions

Type Ia supernova are widely used cosmological probes, limited by residual scatter in measurements. Principal component analysis is a valuable tool that can be used to reduce this scatter and improve cosmological inference, accuracy and efficiency. This work outlines two standardisation processes through principal component analysis derived distance metrics; a photometric standardisation extension and a spectral twinning technique.

This methodology outlines an effective representation of SNIa spectra through PCA compression, capturing almost 90% of the datasets spectral variation in as few as 10 components. The work increases to 15 components capturing fine, feature level variations as well as increased total variation. These components accurately preserve spectral structure, verified through reconstructions and wavelength restriction robustness tests.

With these coefficients, SNIa spectra can efficiently extend lightcurve standardisations, reducing average scatter across the entire dataset by 6%, reducing further to a 7% reduction for a stricter dataset. These improvements are modest but robust and act globally across the dataset. Further analysis also shows potential for reduced datasets to achieve scatter below half of that of the full dataset with only SALT2 photometric standardisations applied.

Using these PCA coefficients, a Euclidean distance was defined, acting as a spectral similarity metric. This metric is validated against full spectral comparisons, demonstrating its ability to track spectral similarity. With this metric, efficient comparisons of SNIa can be conducted, enabling the identification of spectrally similar pairs.

Sufficiently similar pairs can act as "twins" in a local standardisation technique. The closest twin pairs exhibit mean RMS comparisons an order of 10 times lower than that of random SNIa

comparisons. This results in residual scatter between pairs as low as $0.0322\,mag$, providing potential for high precision cosmology. These "twins" are rare pairings, so this method is fundamentally limited by sample size.

Full flux-based spectral comparisons of these "twins" are typically conducted at a high computational cost. Given the rare nature of these pairings, PCA distance was used to provide an efficient twin candidate screening procedure. This enables full flux based twinning on large datasets, with $95 - 100\%$ of twins recovered. Fewer than 10% of non twins can be incorrectly identified under 1.0% closest spectra conditions, drastically reducing dataset size while losing less then 3% of true twins. The thresholds are dynamic allowing for full twin recovery balanced by higher false positive identification.

Supernova cosmology relies on a balance of sample size and precision in standardised SNIa to accurately constrain cosmological parameters. The PCA-based methodology outlined provides two complimentary standardisation approaches addressing this balance. The photometric extension provides a scalable, modest global correction, retaining a large sample size. The PCA-based twinning approach provides a precise local standardisation of rarer spectral twins. Crucially, both approaches exist within the same unified PCA representation, where global standardisation and local spectral matching are defined through a common low-dimensional description of SNIa spectra.

There are several natural extensions to this approach which should be studied. The spectral time series should be factored into this spectral similarity metric, evaluating the evolution of SNIa over time. The reduced dimensionality also facilitates high redshift SNIa with principal components acting as a filter to observational noise, describing intrinsic luminosity characteristics. Methodology improvements should also be considered, including non-linear PCA models. With further refinement, the unified spectral similarity framework proposed offers a promising pathway toward higher precision cosmological constraints, advancing the use of Type Ia supernovae as measures of Universe expansion.

References

[1] Hubble E. A relation between distance and radial velocity among extra-galactic nebulae. Proceedings of the National Academy of Sciences. 1929;15(3):168-73. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.15.3.168.

[2] Hogg DW. Distance measures in cosmology; 2000. Available from: https://arxiv.org/abs/astro-ph/9905116.

[3] Carroll BW, Ostlie DA. An Introduction to Modern Astrophysics. 2nd ed. San Francisco, CA: Addison-Wesley; 2007.

[4] Tornambé A, Piersanti L. Clues on Type Ia Supernovae Progenitors. In: Turatto M, Benetti S, Zampieri L, Shea W, editors. 1604-2004: Supernovae as Cosmological Lighthouses. vol. 342 of Astronomical Society of the Pacific Conference Series; 2005. p. 169.

[5] Riess AG, Filippenko AV, Challis P, Clocchiatti A, Diercks A, Garnavich PM, et al. Observational Evidence for an Accelerating Universe and a Cosmological Constant. The Astronomical Journal. 1998;116(3):1009-38.

[6] Perlmutter S, Aldering G, Goldhaber G, Knop RA, Nugent P, Castro PG, et al. Measurements of Ω and Λ from 42 High-Redshift Supernovae. The Astrophysical Journal. 1999 Jun;517(2):565–586. Available from: http://dx.doi.org/10.1086/307221.

[7] Dhawan S, Goobar A, Smith M, Johansson J, Rigault M, Nordin J, et al. The Zwicky Transient Facility Type Ia supernova survey: first data release and results. . 2022 Feb;510(2):2228-41.

[8] Brout D, Sako M, Scolnic D, Kessler R, D'Andrea CB, Davis TM, et al. First Cosmology Results Using Type Ia Supernovae from the Dark Energy Survey: Photometric Pipeline and Light-curve Data Release. . 2019 Mar;874(1):106.

[9] Rest A, Scolnic D, survey PS. Cosmological Constraints Measurements of Type Ia Supernovae Discovered during the first 1.5 Years of the Pan-STARRS1 Survey. In: American Astronomical Society Meeting Abstracts #223. vol. 223 of American Astronomical Society Meeting Abstracts; 2014. p. 316.07.

[10] Bazin G, Ruhlmann-Kleider V, Palanque-Delabrouille N, Rich J, Aubourg E, Astier P, et al. Photometric selection of Type Ia supernovae in the Supernova Legacy Survey. . 2011 Oct;534:A43.

[11] Rigault M, Smith M, Goobar A, Maguire K, Dimitriadis G, Burgaz U, et al.. ZTF SN Ia DR2: Overview; 2024. Available from: https://arxiv.org/abs/2409.04346.

[12] Fakhouri HK, Boone K, Aldering G, Antilogus P, Aragon C, Bailey S, et al. Improving Cosmological Distance Measurements Using Twin Type Ia Supernovae. . 2015 Dec;815(1):58.

[13] Bailey S, Aldering G, Antilogus P, Aragon C, Baltay C, Bongard S, et al. Using spectral flux ratios to standardize SN Ia luminosities. . 2009 Jun;500(2):L17-20.

[14] Saunders C, Aldering G, Antilogus P, Bailey S, Baltay C, Barbary K, et al. SNEMO: Improved Empirical Models for Type Ia Supernovae. . 2018 Dec;869(2):167.

[15] Chandrasekhar S. The Maximum Mass of Ideal White Dwarfs. . 1931 Jul;74:81.

[16] Colgate SA, McKee C. Early Supernova Luminosity. . 1969 Aug;157:623.

[17] Anderson JP. A meta-analysis of core-collapse supernova ^{56}Ni masses. . 2019 Aug;628:A7.

[18] Phillips MM, Lira P, Suntzeff NB, Schommer RA, Hamuy M, Maza J. The Reddening-Free Decline Rate Versus Luminosity Relationship for Type IA Supernovae. . 1999 Oct;118(4):1766-76.

[19] Rigault M, Copin Y, Aldering G, Antilogus P, Aragon C, Bailey S, et al. Evidence of environmental dependencies of Type Ia supernovae from the Nearby Supernova Factory indicated by local Hα. . 2013 Dec;560:A66.

[20] Livio M, Mazzali P. On the progenitors of Type Ia supernovae. . 2018 Mar;736:1-23.

[21] Rybicki GB, Lightman AP. Radiative Processes in Astrophysics; 1986.

[22] Filippenko AV. Optical Spectra of Supernovae. . 1997 Jan;35:309-55.

[23] Branch D, Dang LC, Hall N, Ketchum W, Melakayil M, Parrent J, et al. Comparative Direct Analysis of Type Ia Supernova Spectra. II. Maximum Light. . 2006 Apr;118(842):560-71.

[24] Hachinger S, Mazzali PA, Tanaka M, Hillebrandt W, Benetti S. Spectral luminosity indicators in Type Ia supernovae. Understanding the (SiII) line-strength ratio and beyond. . 2008 Sep;389(3):1087-96.

[25] Gewers FL, Ferreira GR, de Arruda HF, Silva FN, Comin CH, Amancio DR, et al. Principal Component Analysis: A Natural Approach to Data Exploration. arXiv e-prints. 2018 Apr:arXiv:1804.02502.

[26] Mahalanobis PC. On the Generalised Distance in Statistics. Sankhya A. 1936. Reprinted in Sankhya A, 80(1), 1–7 (2018).

[27] Astropy Collaboration, Robitaille TP, Tollerud EJ, Greenfield P, et al. Astropy: A community Python package for astronomy. Astronomy & Astrophysics. 2013;558:A33.

[28] Astropy Collaboration, Price-Whelan AM, Sipőcz BM, Günther HM, et al. The Astropy Project: Building an open-science project and status of the v2.0 core package. Astronomy & Astrophysics. 2018;616:A1.

[29] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011 Oct;12:2825-30.

[30] Guy J, Astier P, Baumont S, Hardin D, Pain R, Regnault N, et al. SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators. . 2007 Apr;466(1):11-21.

[31] Tripp R. A two-parameter luminosity correction for Type IA supernovae. . 1998 Mar;331:815-20.

[32] James JB, Davis TM, Schmidt BP, Kim AG. Spectral diversity of Type Ia supernovae. . 2006 Aug;370(2):933-40.

[33] Nugent P, Phillips M, Baron E, Branch D, Hauschildt P. Evidence for a Spectroscopic Sequence among Type 1a Supernovae. . 1995 Dec;455:L147.

[34] Foley RJ, Kasen D. Measuring Ejecta Velocity Improves Type Ia Supernova Distances. . 2011 Mar;729(1):55.

[35] Branch D, Baron E, Hall N, Melakayil M, Parrent J. Comparative Direct Analysis of Type Ia Supernova Spectra. I. SN 1994D. . 2005 Jun;117(832):545-52.

[36] Kelly PL, Hicken M, Burke DL, Mandel KS, Kirshner RP. Hubble Residuals of Nearby Type Ia Supernovae are Correlated with Host Galaxy Masses. . 2010 Jun;715(2):743-56.

[37] Childress M, Aldering G, Antilogus P, Aragon C, Bailey S, Baltay C, et al. Host Galaxy Properties and Hubble Residuals of Type Ia Supernovae from the Nearby Supernova Factory. . 2013 Jun;770(2):108.

[38] Hinton G, Salakhutdinov R, Rachmad Y. Reducing the Dimensionality of Data with Neural Networks. Science (New York, NY). 2006 08;313:504-7.

[39] Schölkopf B, Smola A, Müller KR. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation. 1998 07;10:1299-319.

[40] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning; 2006.

A Data Sample

A.1 Spectroscopic File Naming Convention

Spectroscopic files take the form `ZTFname_Date_Instrument_Index.ascii`. For example spectrum file `ZTF18aabtiph_20180309_SEDm_0.ascii` is a spectrum taken of a Type Ia supernova (SNIa) with ztfname ZTF18aabtiph using the Spectral Energy Distribution Machine (see table 4 for full instrument code conversion) on 9 March 2018. This index is not representative of when the spectrum was taken and is strictly a naming convention to allow for multiple spectra per SNIa. An index of 0 is the first available spectrum in the dataset.

A.2 Spectroscopic Instruments

Table 4: Spectroscopic instrument identifiers extracted from filename metadata and their corresponding telescope or instrument names.

Code	Telescope / Instrument
SEDm	Spectral Energy Distribution Machine (P60)
P200	Palomar 200-inch (Hale Telescope, DBSP)
Lick-3m	Shane 3m Telescope (Lick Observatory)
Lijiang-2.4m	Lijiang Observatory 2.4m Telescope
NTT	New Technology Telescope (ESO La Silla)
Xinglong-2.2m	Xinglong Observatory 2.2m Telescope
LCO-FLOYDS	FLOYDS Spectrograph (Las Cumbres Observatory)
APO	Apache Point Observatory 3.5m Telescope
SALT	Southern African Large Telescope
MDM-2.4	MDM Observatory 2.4m Telescope
LT	Liverpool Telescope
NOT	Nordic Optical Telescope
Ekar	Asiago 1.82m (Ekar Telescope)
DCT	Discovery Channel Telescope
GTC	Gran Telescopio Canarias
Keck	W. M. Keck Observatory
Plaskett	Dominion Astrophysical Observatory (Plaskett Telescope)
SOAR	Southern Astrophysical Research Telescope
SNIFS	SuperNova Integral Field Spectrograph (UH 2.2m)
Bok	Bok 2.3m Telescope
Gemini-N	Gemini North Telescope
MMT	Multiple Mirror Telescope
LCO-duPont	du Pont Telescope (Las Campanas Observatory)
Other	Unspecified
WHT	William Herschel Telescope
Kanata	Kanata Telescope
Magellan-Baade	Magellan Baade Telescope
Magellan-Clay	Magellan Clay Telescope
FLWO-1.5m	Fred Lawrence Whipple Observatory 1.5m Telescope
Gemini-S	Gemini South Telescope
Lick	Lick Observatory (unspecified instrument)
TNG	Telescopio Nazionale Galileo
VLT-FORS	Very Large Telescope (FORS instrument)
IRTF	Infrared Telescope Facility
CMO-2.5m	Caucasus Mountain Observatory 2.5m Telescope

B SALT2 PCA Extension

B.1 PC Coefficients

Below are tables with containing the principal component coefficients for varying methods.

- (1) Standardisation in form

$$\mu = m_B - M + \alpha x_1 - \beta c$$

Coefficient	Value
α	0.123
β	2.932

Table 5: SALT2 standardisation coefficients, resulting in a standardisation with scatter of 0.2400 mag.

- (2) Standardisation in form

$$\mu = m_B - M + \alpha x_1 - \beta c + \sum_i \gamma_i p_i$$

where SALT2 parameters α and β are fixed first then PC coefficients determined.

Coefficient	Value
α	0.123
β	2.932
PC1	0.0167
PC2	-0.0452
PC3	0.0401
PC4	-0.0148
PC5	0.0012
PC6	0.0388
PC7	-0.0294
PC8	0.0014
PC9	0.0475
PC10	0.0248
PC11	-0.0071
PC12	0.0278
PC13	-0.0346
PC14	0.0403
PC15	-0.0171

Table 6: SALT2 + PCA standardisation coefficients, resulting in a standardisation with scatter of 0.2104 mag.

- (3) Standardisation in form

$$\mu = m_B - M + \alpha x_1 - \beta c + \sum_i \gamma_i p_i$$

where SALT2 parameters α and β are free and fit simultaneously with PC coefficients.

Coefficient	Value
α	0.077
β	2.858
PC1	-0.0002
PC2	0.0389
PC3	-0.0451
PC4	0.0277
PC5	0.0161
PC6	-0.0679
PC7	0.0219
PC8	0.0095
PC9	-0.0443
PC10	-0.0272
PC11	0.0089
PC12	-0.0319
PC13	0.0404
PC14	-0.0453
PC15	0.0178

Table 7: SALT2 + PCA standardisation coefficients, resulting in a standardisation with scatter of 0.2078 mag.

- (4) Standardisation in form

$$\mu = m_B - M + \sum_i \gamma_i p_i$$

Coefficient	Value
PC1	0.2445
PC2	-0.0803
PC3	-0.1619
PC4	-0.1219
PC5	-0.0826
PC6	-0.1613
PC7	0.0456
PC8	0.0507
PC9	-0.0624
PC10	-0.0340
PC11	0.0646
PC12	-0.0859
PC13	0.0463
PC14	-0.1167
PC15	0.0054

Table 8: PCA standardisation coefficients, resulting in a standardisation with scatter of 0.4291 mag.

B.2 Number of PCs Used

Table 9 quantifies the residual scatter as more PCs are introduced to the SALT2 standardisation extension. Figure 17 illustrates this scatter reduction across varying clippings, including the tight 1σ cut provides significant reduction but drastically reduces sample size.

PCs Used	Scatter /mag	Scatter Reduction (relative to SALT2) /mag
1	0.2394	0.0006
2	0.2351	0.0049
3	0.2317	0.0083
4	0.2312	0.0088
5	0.2312	0.0088
6	0.2279	0.0121
7	0.2260	0.0140
8	0.2260	0.0140
9	0.2209	0.0191
10	0.2196	0.0205
11	0.2194	0.0206
12	0.2177	0.0223
13	0.2149	0.0251
14	0.2111	0.0289
15	0.2104	0.0296

Table 9: Hubble residual scatter as a function of the number of principal components included in the PCA-extended SALT2 standardisation. The scatter is defined as the standard deviation of residuals relative to a Flat ΛCDM cosmology. The scatter reduction is measured relative to the SALT2 only baseline, demonstrating progressive improvement with increasing spectral information.

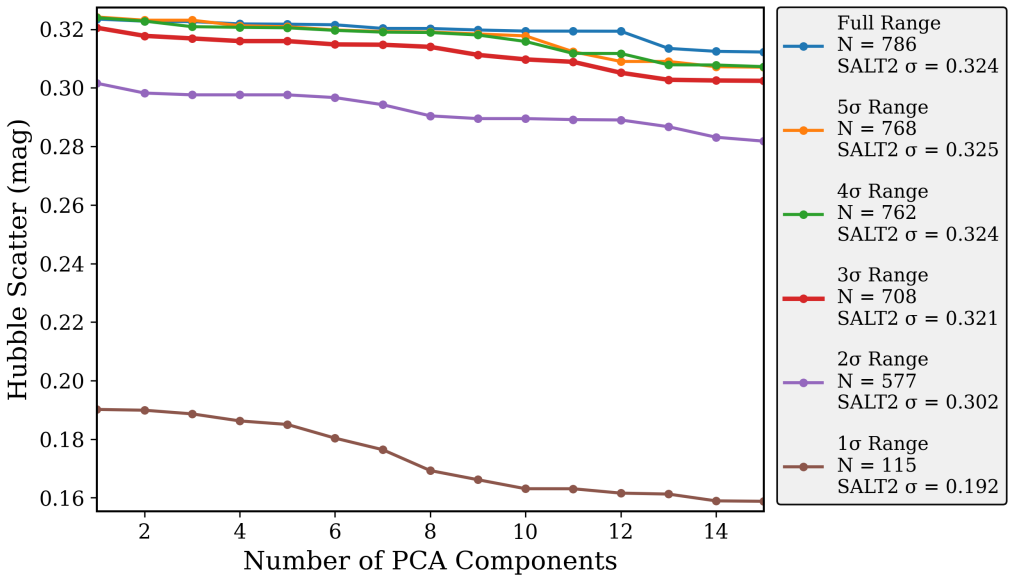

Figure 17: Hubble scatter of Type Ia supernovae as a function of the number of PCA components used to correct SALT2 distance moduli. Lines correspond to different standard deviation clipping levels applied to the PCA coefficients, SNIa with any coefficient outside of these ranges are removed. The scatter is the standard deviation of residuals relative to the theoretical distance modulus from a Flat ΛCDM cosmology.

B.3 Cosmological Constraints

A bootstrap resampling was applied to estimate the uncertainty in Ω_m values. The dataset was repeatedly resampled and Ω_m was refit for both SALT2 and SALT2 with PCA extension. The resulting distributions shown in figure 18 represent the statistical spread in the cosmological parameter.

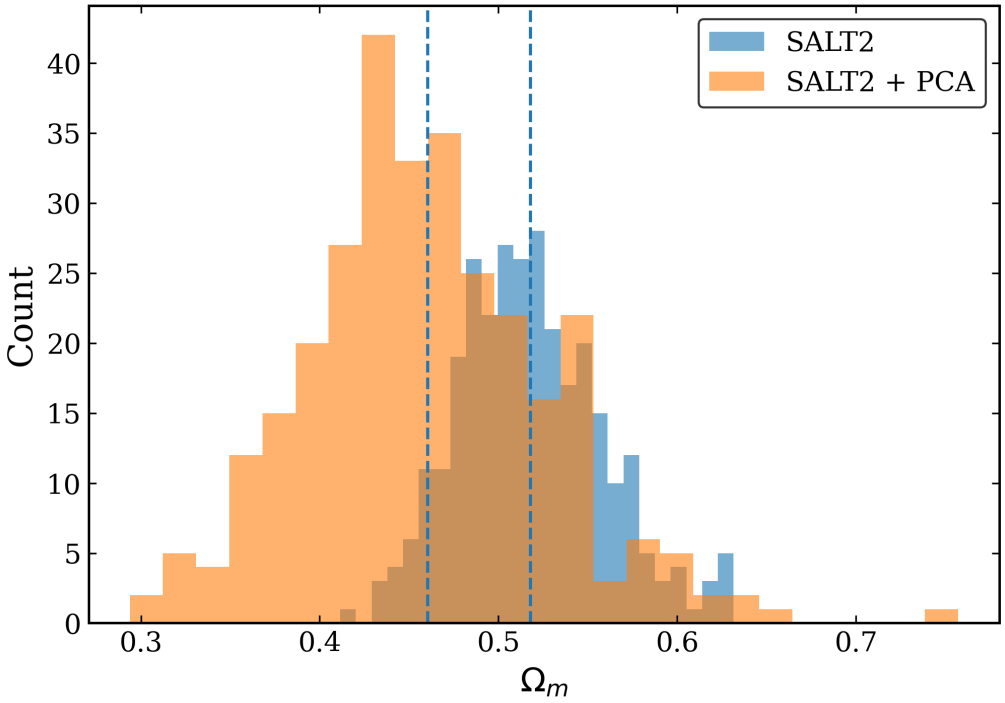

Figure 18: Determined Ω_m values using SALT2 and PCA extended SALT2 standardisations. 15PCs are used for the extended standardisation.

C Spectral Twinning in PCA Space

Figure 19 illustrated the number of PCs used to identify twins with the lowest Hubble residuals.

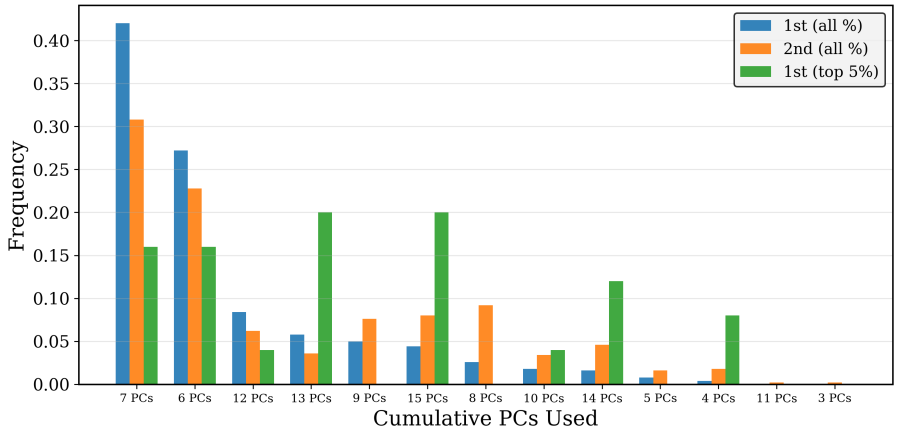

Figure 19: Results for the number of cumulative PCs used for twinning which achieve the lowest Hubble residuals. A 5 fold validation was conducted to reduce the influence of potential outliers at low percentiles.

C.1 Higher Redshift Spectral Twinning in PCA Space

Figure 20 illustrates an approach used to twin higher redshift SNIa outside of the applied cuts to a low redshift SNIa used to train the PCA model.

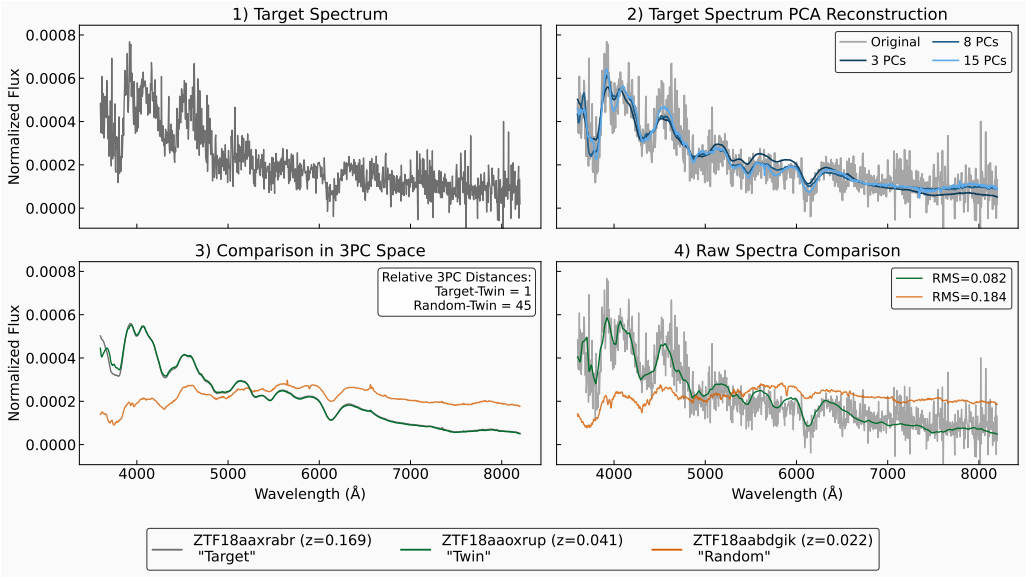

Figure 20: An illustration of a higher redshift spectral twinning method. The target spectrum (1) is deconstructed via PCA (2) trained on low redshift SNIa, this deconstructed spectra is then compared to the PCA training dataset SNIa using the first 3 PCs (3). The target SNIa can then be 'twinned' to a low redshift SNIa and their intrinsic brightnesses can be assumed to be the same (or further corrections implimented). (4) shows that the twinning in 3 PC space identifies spectra with similar spectral shape despite having much lower signal to noise.

D Robustness Tests

D.1 Wavelength Restriction

To evaluate the robustness of PCA representation of spectra and the physical spectral trends, the spectral decomposition is limited to reduced wavelength regions. The unseen regions of the spectrum are extrapolated and compared with the true spectrum. This determines whether overall physical trends are captured through PCA rather than just pattern simplification. Figure 21 illustrates this wavelength restriction and reconstruction, figure 22 shows the derivative flux values assessing detail reconstruction ability.

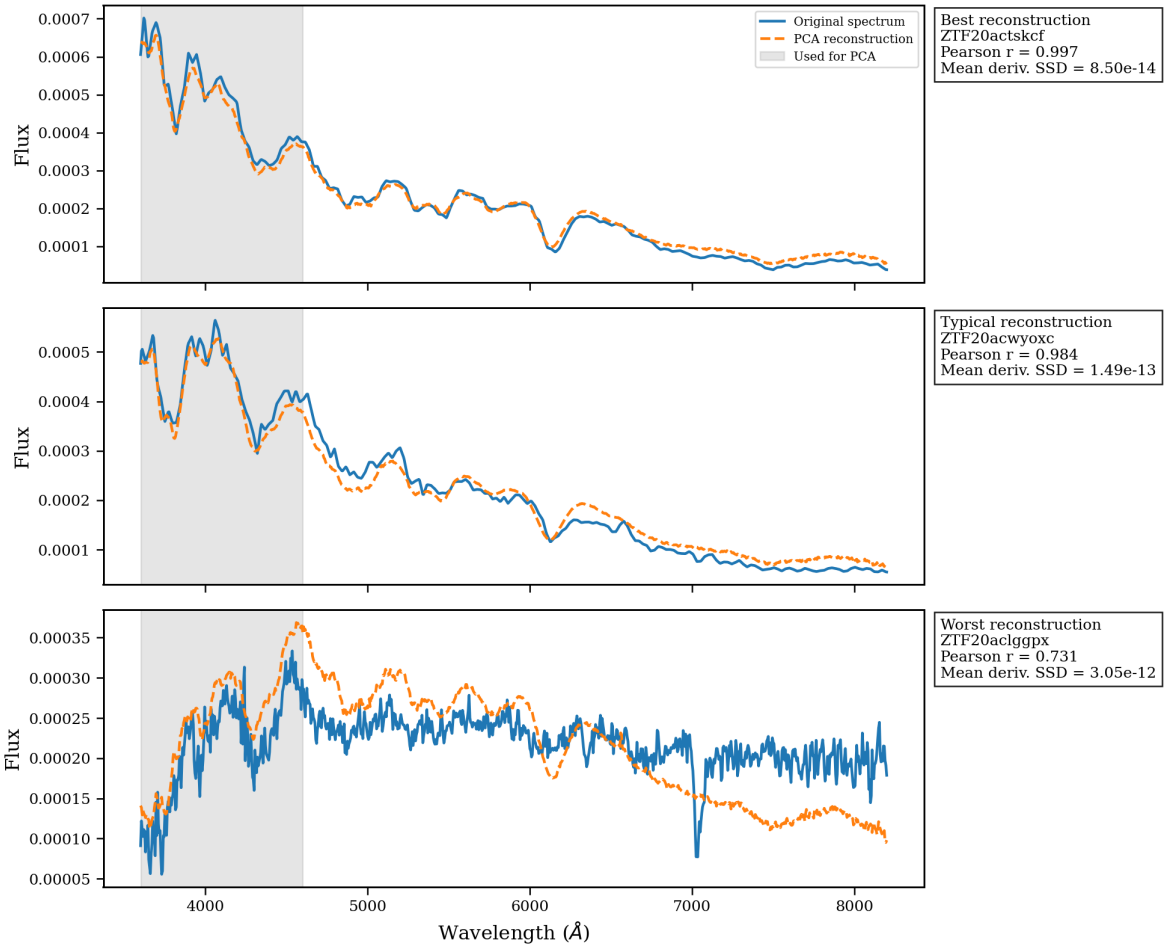

Figure 21: Representative examples of PCA reconstruction of SNIa spectra using a restricted wavelength range (shaded region). The original spectra (solid) are compared to PCA reconstructions (dashed) extrapolated beyond the observed region. The top panel shows a good, middle a typical and bottom a poor reconstruction, selected based on hidden-region Pearson correlation and derivative error metrics. While PCA accurately recovers both continuum shape and spectral features in most cases, deviations in the worst example highlight limitations when reconstructing more complex or noisy spectra.

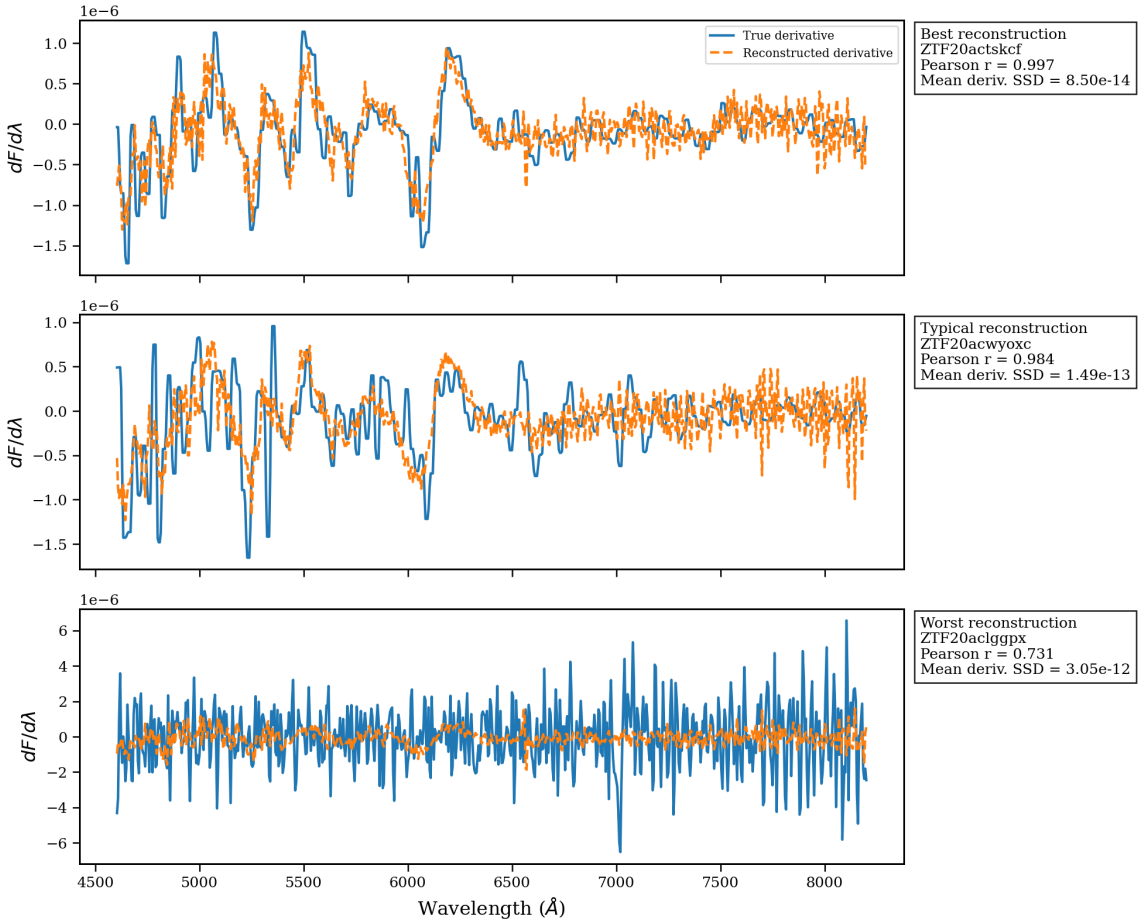

Figure 22: Comparison of spectral derivatives in the hidden (unobserved) wavelength region for the same examples shown in Figure 21. The true derivatives (solid) and PCA reconstructions (dashed) demonstrate that PCA captures both the global spectral shape and finer features. The degradation in the worst case example illustrates reduced ability to reconstruct detailed spectral features.

Reconstruction performance is dependent on the 'seen' wavelength range. Figures 23 & 24 quantify this performance across varying 'seen' regions.

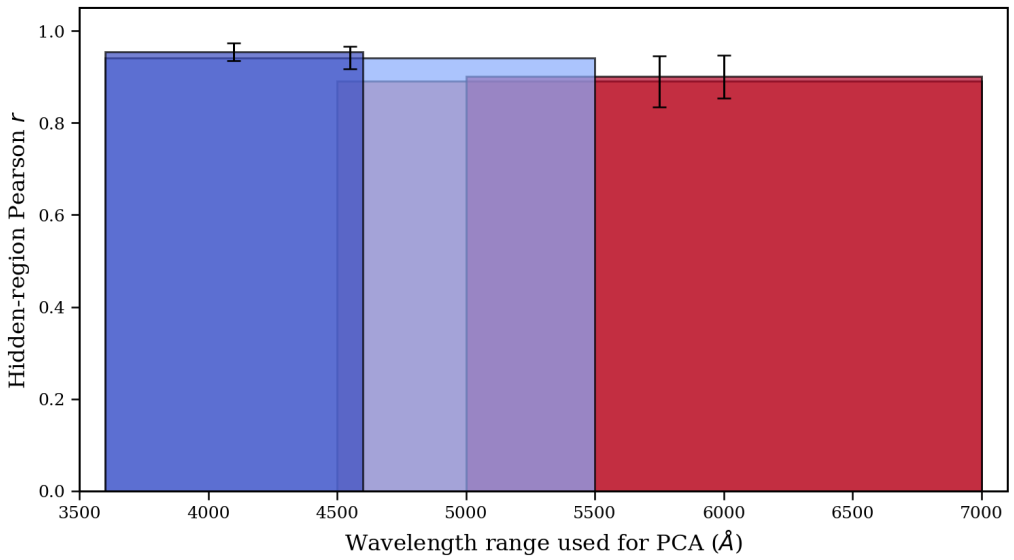

Figure 23: Mean Pearson r coefficient values between the true and PCA-reconstructed hidden regions of the spectra. Higher values indicate a closer correlation between the spectra and a better extrapolation of the spectra. The four seen regions are illustrated, showing that all have high r values. PCA-reconstruction models achieve higher values when bluer regions are seen. 15 SNIa are used.

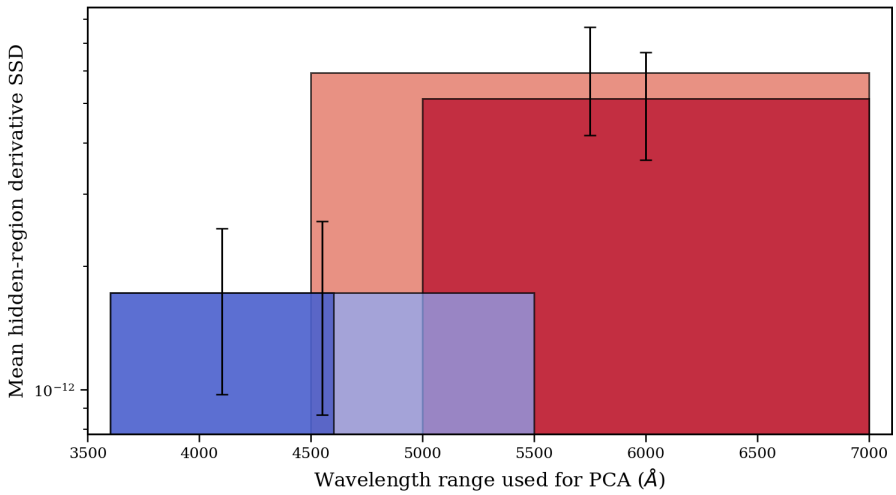

Figure 24: Mean hidden-region derivative squared spectral distance (SSD) for PCA reconstructions using different wavelength intervals. The PCA-reconstruction uses the seen regions shown and the SSD is evaluated on the complementary, hidden region. Lower mean SSD across the 15 SNIa used indicated higher similarity between the true and reconstructed spectrum. As shown, bluer seen wavelengths achieve lower SSD.

D.2 Simulated Noise

Robustness tests were performed on the spectral twinning method by applying perturbations to the spectra to assess the stability of PCA-based twin SNIa pairs under noise. Several noise models are applied:

- Gaussian: Additive, wavelength-independent random noise scaled to the spectrum variance, representative of detector noise.
- Poisson Like: Flux-dependent noise with variance proportional to $\sqrt{\lambda}$ approximating photon-counting behavior.
- Correlated: Smooth, wavelength-correlated perturbations generated via Gaussian filtering mimicking calibration systematics.
- Tilt: Multiplicative linear gradient applied across the full wavelength range, representing calibration errors.
- Chunk: Additive offsets applied to wavelength segments simulating local calibration errors.

Figure 25 shows how the 'best twin' changes under noise injection and Figures 26, 27, 28 show the retention rate of this 'best twin' being within the top 1%, 2.5% and 5% under noise injection respectively. Figures 29, 30, 31 indicate the number of the top 1%, 2.5%, 5% closest matches remaining inside these top percentiles under noise. Figure 32 assesses the shifting of twin positions under noise quantified through Spearman rank and figure 33 indicates the average shift in rank of the SNIa under noise.

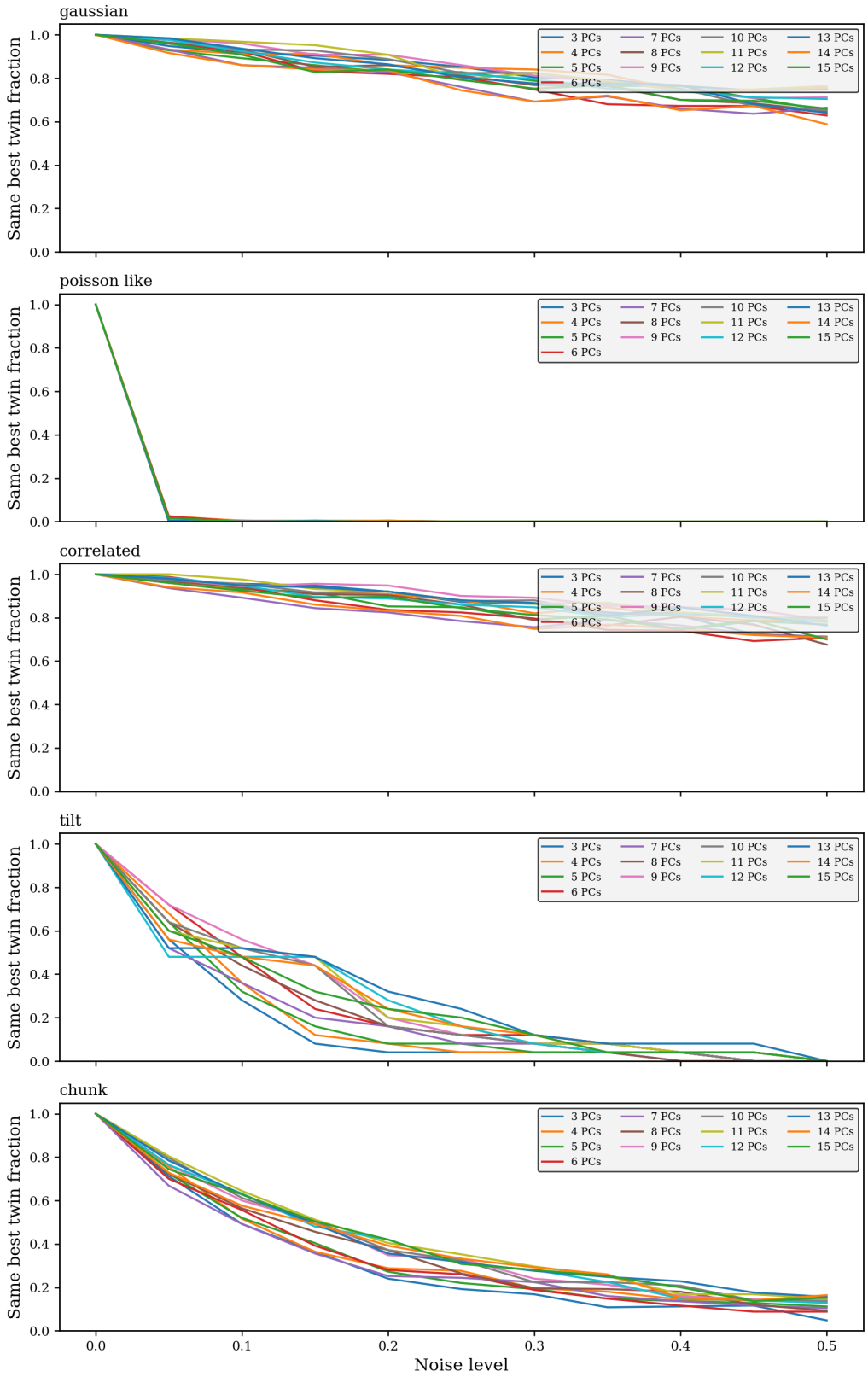

Figure 25: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of cases in which the nearest neighbor (best twin) remains unchanged after noise is applied. Values are averaged over 25 supernovae.

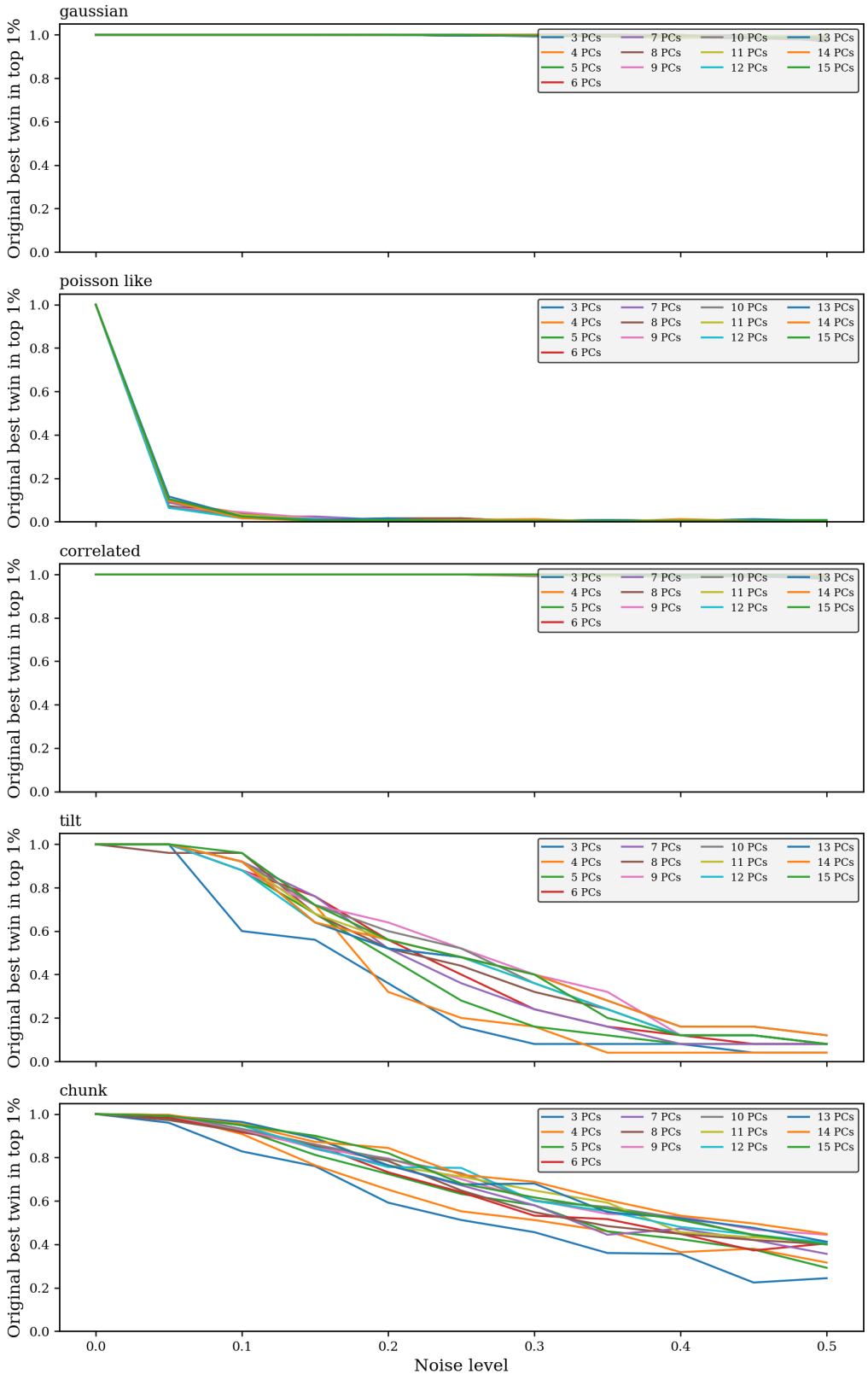

Figure 26: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of cases in which the original best twin remains within the top 1% of nearest neighbors after noise is applied. Values are averaged over 25 supernovae.

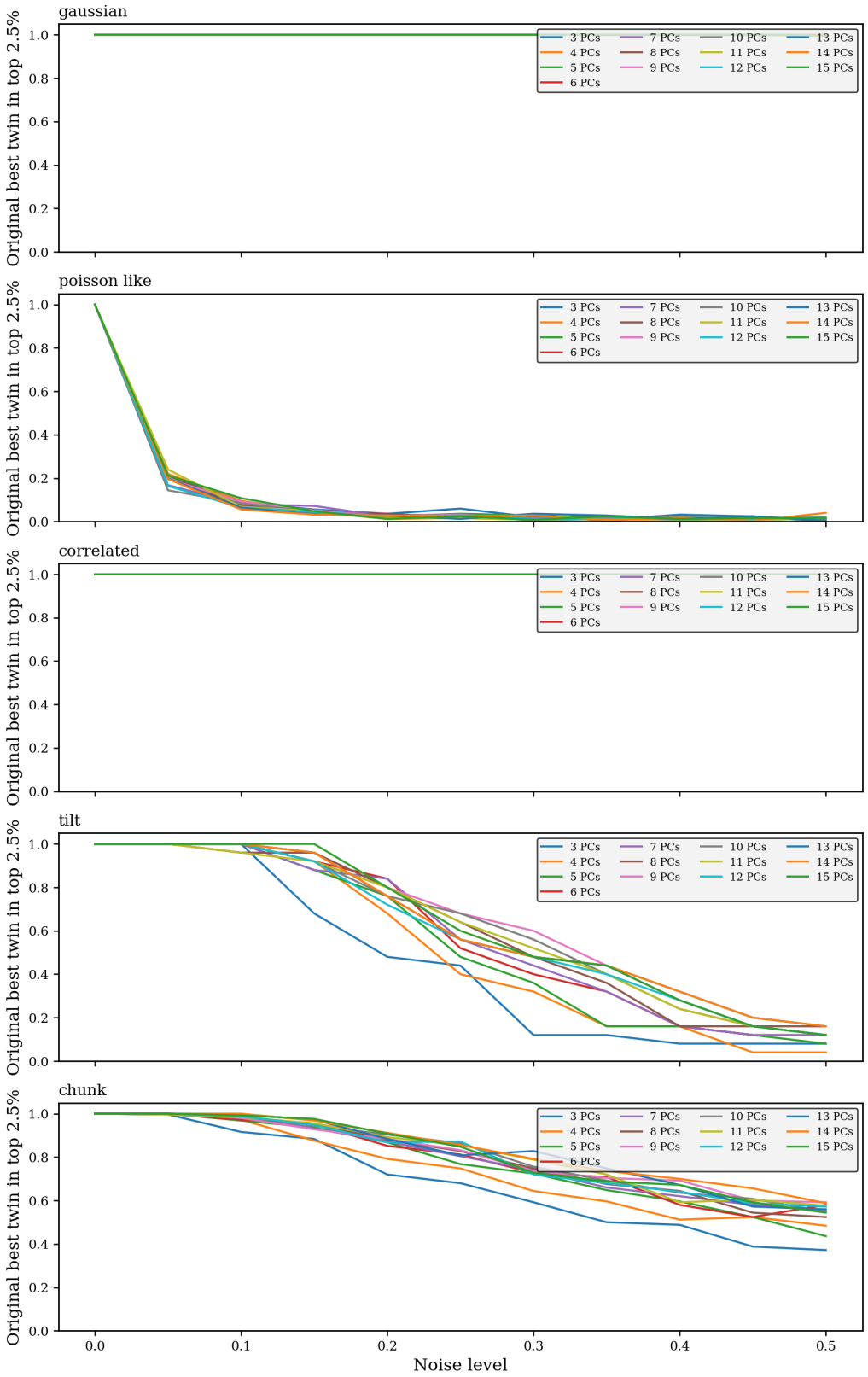

Figure 27: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of cases in which the original best twin remains within the top 2.5% of nearest neighbors after noise is applied. Values are averaged over 25 supernovae.

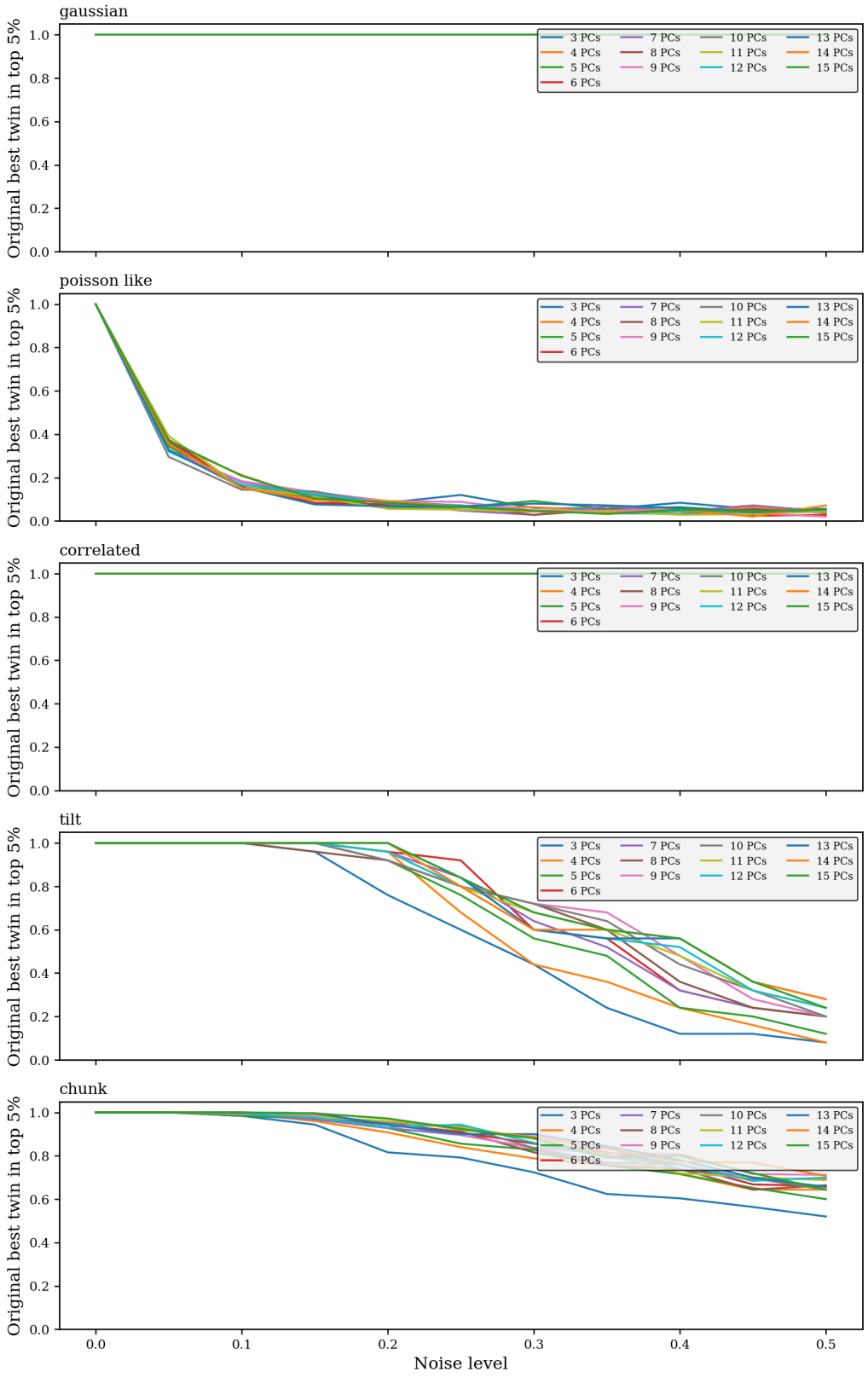

Figure 28: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of cases in which the original best twin remains within the top 5% of nearest neighbors after noise is applied. Values are averaged over 25 supernovae.

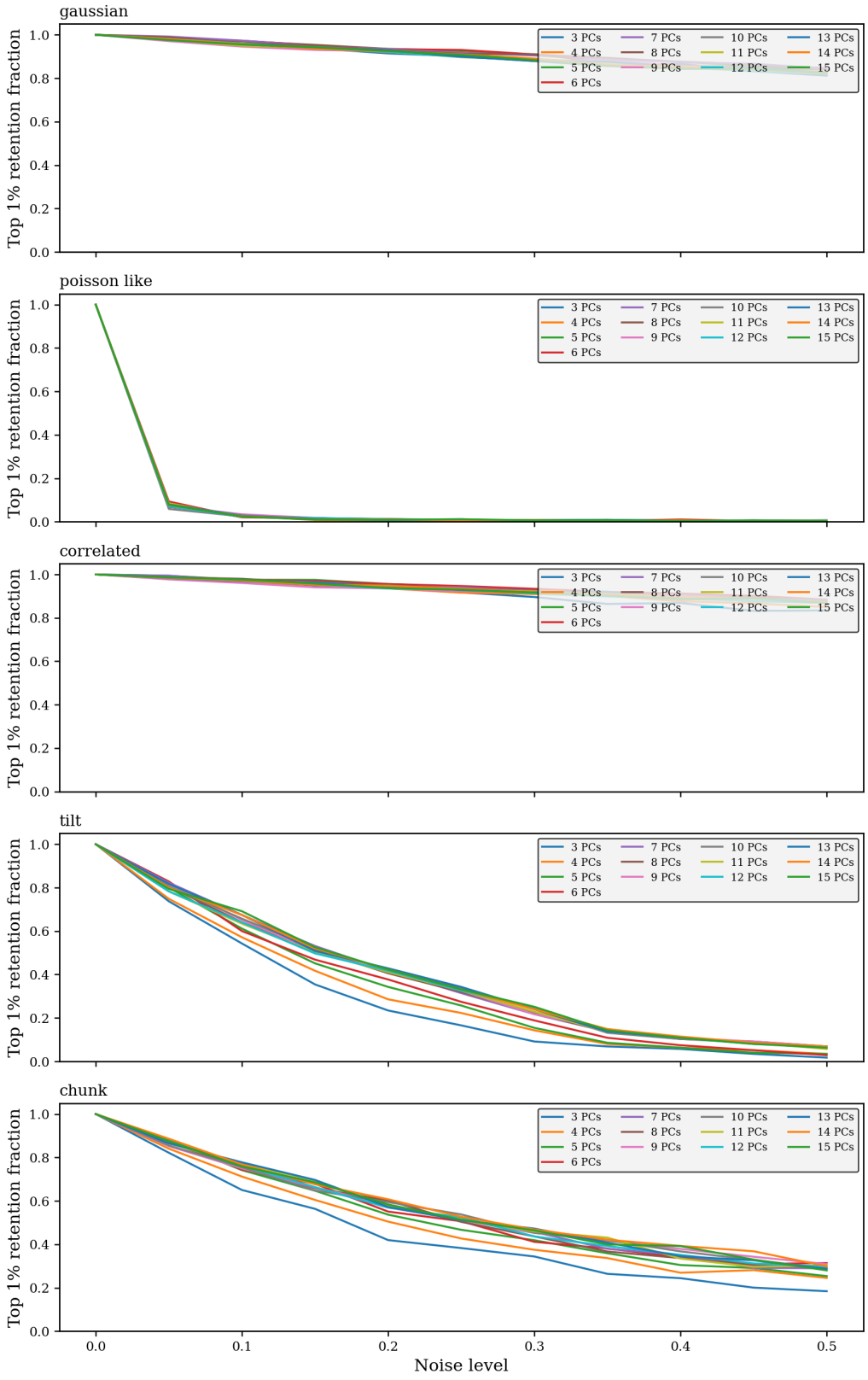

Figure 29: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of the original top 1% nearest neighbors retained after noise is applied. Values are averaged over 25 supernovae.

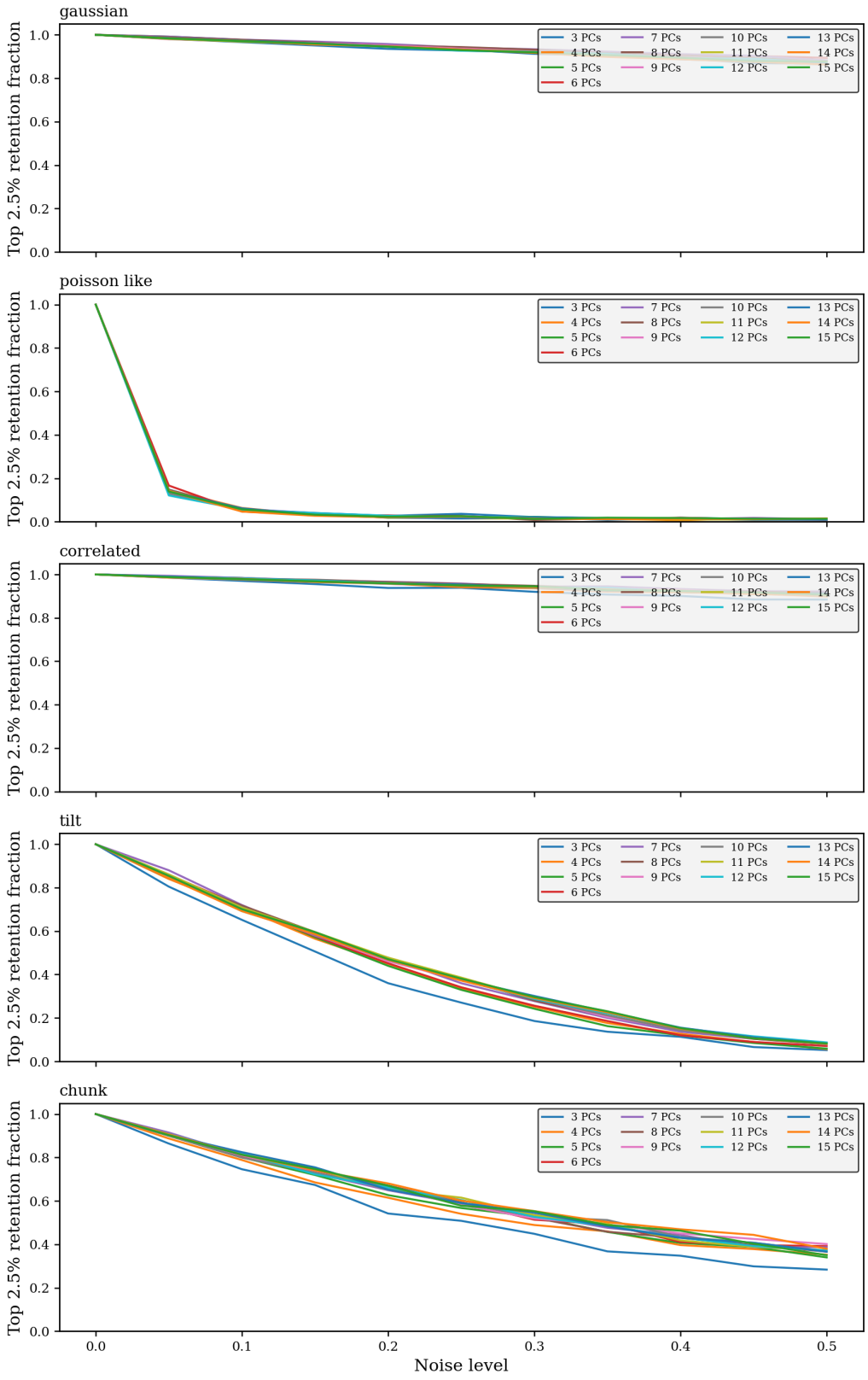

Figure 30: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of the original top 2.5% nearest neighbors retained after noise is applied. Values are averaged over 25 supernovae.

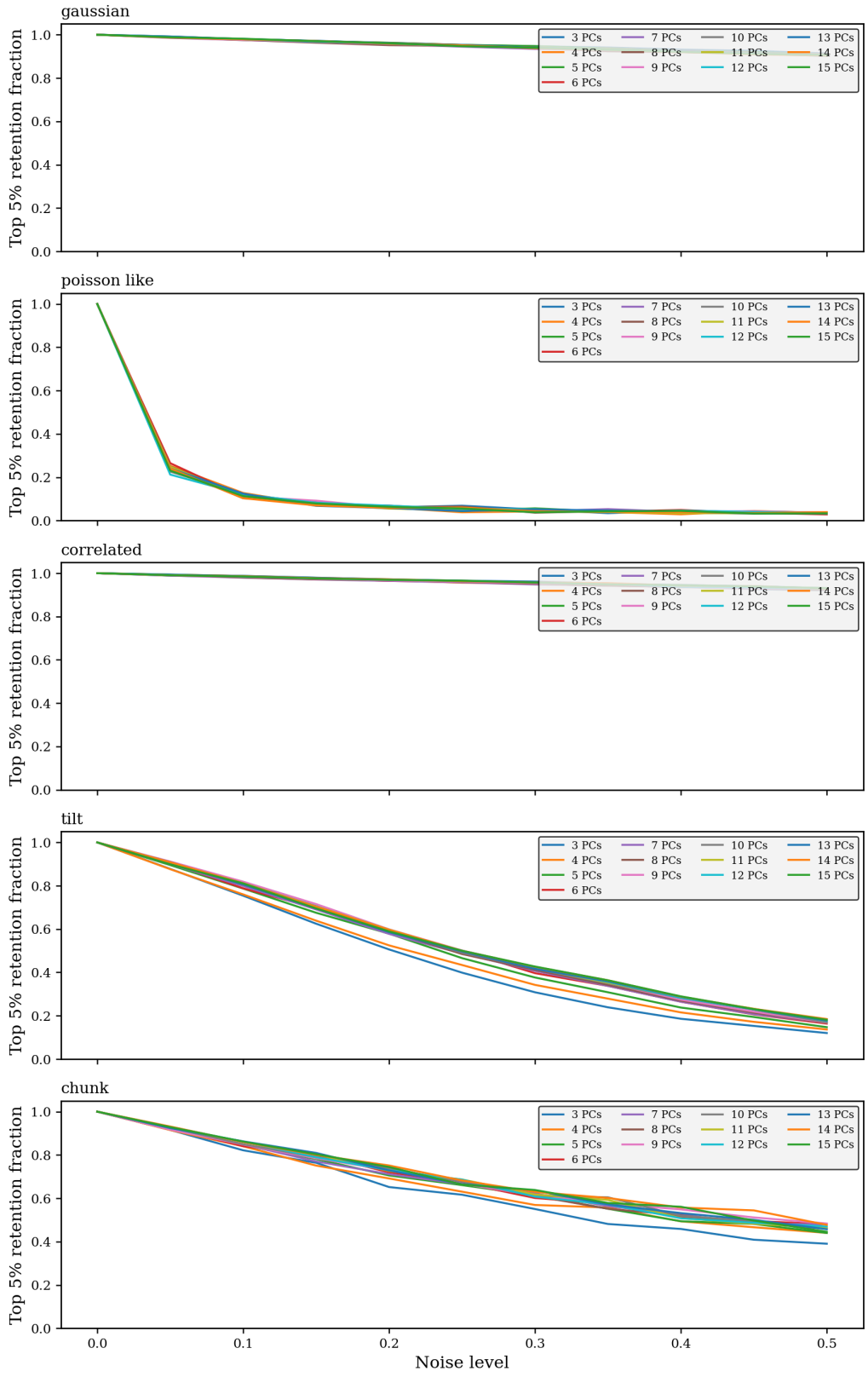

Figure 31: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the fraction of the original top 5% nearest neighbors retained after noise is applied. Values are averaged over 25 supernovae.

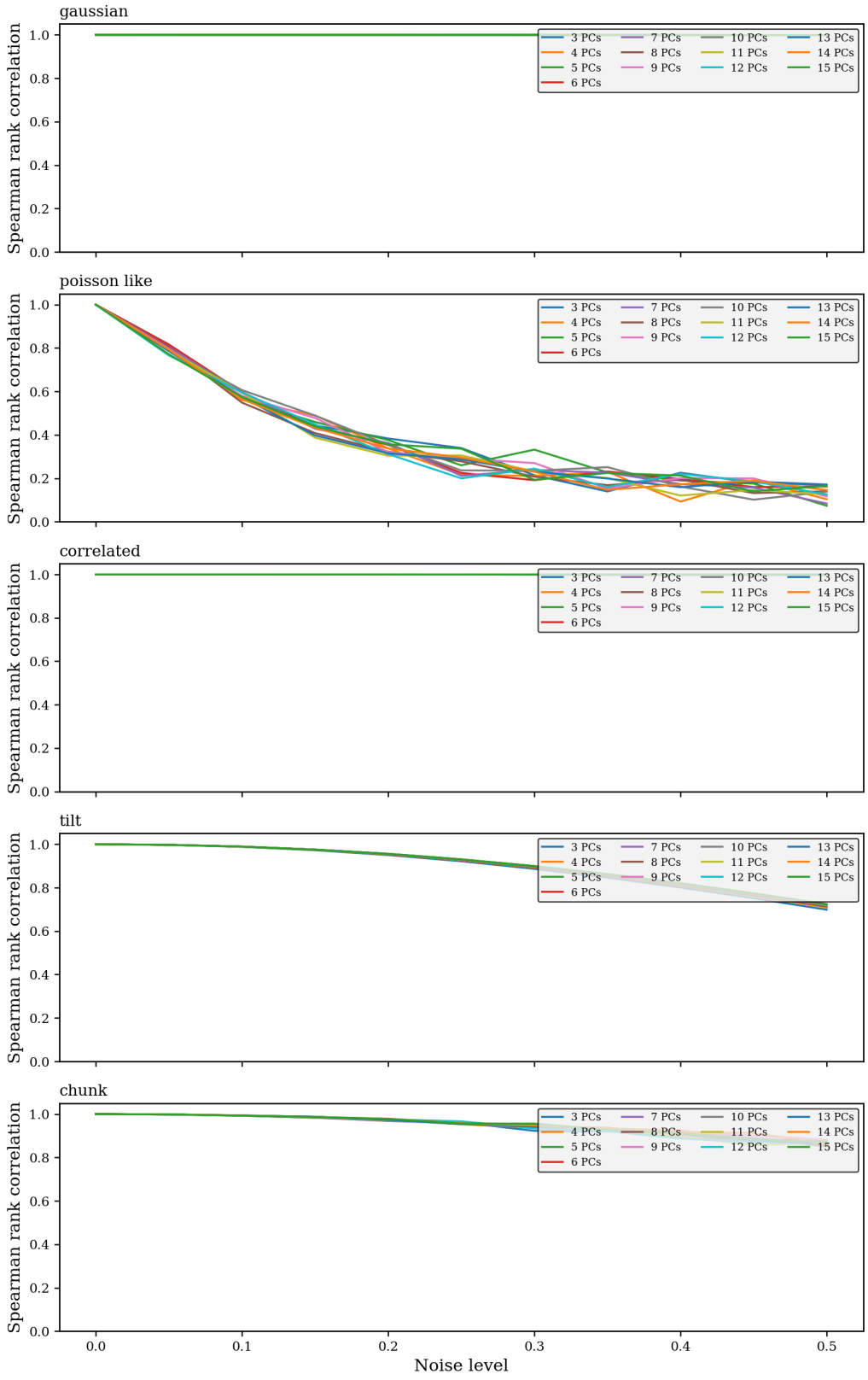

Figure 32: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the Spearman rank correlation between the full pairwise distance rankings before and after noise is applied. Values are averaged over 25 supernovae.

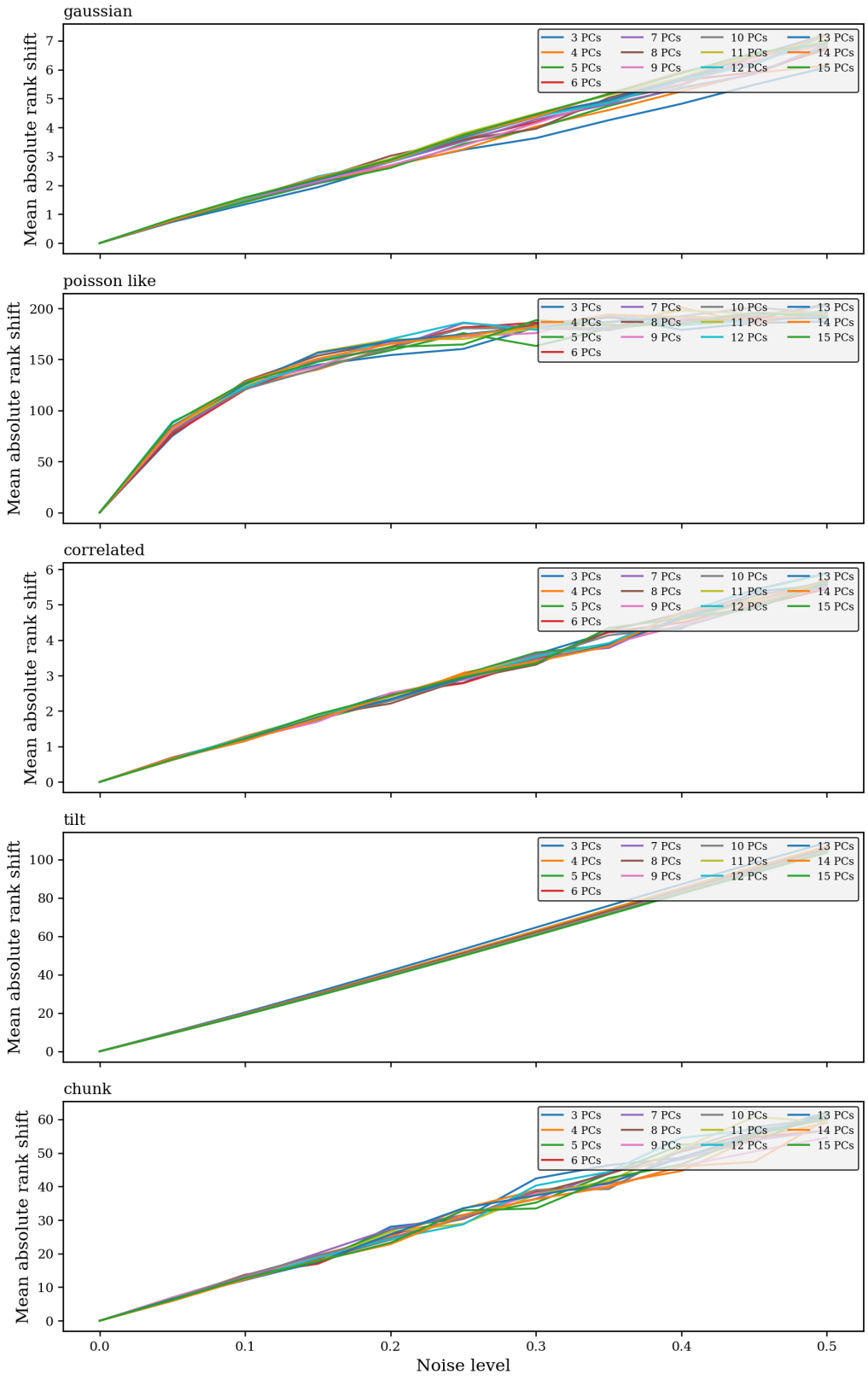

Figure 33: Robustness of PCA-based twin identification under simulated noise perturbations. Each panel shows the performance of the method as a function of noise level for different principal component (PC) truncations (3–15 PCs), for a given noise model (Gaussian, Poisson-like, correlated, tilt, and chunk). The noise level indicates the strength of the applied noise. The metric shown is the mean absolute change in rank position of supernovae in the nearest-neighbor ordering after noise is applied. Values are averaged over 25 supernovae.