On the Automated and Objective Detection of Emission Lines in Faint-Object Spectroscopy

Sungryong Hong,1 Arjun Dey,1,2 and Moore K. M. Prescott3

Received 2014 February 10; accepted 2014 September 24; published 2014 October 24

ABSTRACT. Modern spectroscopic surveys produce large spectroscopic databases, generally with sizes well beyond the scope of manual investigation. The need arises, therefore, for an automated line detection method with objective indicators for detection significance. In this paper, we present an automated and objective method for emission line detection in spectroscopic surveys and apply this technique to observed spectra from a Lyα emitter survey at \( z \sim 2.7 \), obtained with the Hectospec spectrograph on the MMT Observatory (MMTO). The basic idea is to generate on-source (signal plus noise) and off-source (noise only) mock observations using Monte Carlo simulations, and calculate completeness and reliability values, \((C, R)\), for each simulated signal. By comparing the detections from real data with the Monte Carlo results, we assign the completeness and reliability values to each real detection. From 1574 spectra, we obtain 881 raw detections and, by removing low reliability detections, we finalize 652 detections from an automated pipeline. Most of high completeness and reliability detections, \((C, R) \approx (1.0, 1.0)\), are robust detections when visually inspected; the low \( C \) and \( R \) detections are also marginal on visual inspection. This method of detecting faint sources is dependent on the accuracy of the sky subtraction.

Online material: color figures

1. INTRODUCTION

Modern spectroscopic surveys produce large and relatively uniform spectroscopic databases (e.g., AGES, BOSS, BigBOSS, and SEGUE; Kochanek et al. 2012; Dawson et al. 2013; Schlegel et al. 2009; Yanny et al. 2009). Though many preprocessing reduction pipelines exist, the reduced spectra traditionally have been analyzed visually. Since the advent of the Sloan Digital Sky Survey (SDSS; York et al. 2000; Ahn et al. 2012), we have now entered an era of immense spectroscopic databases that defy manual inspection and necessitate automation. For example, Bolton et al. (2012) presented automated pipelines to classify galaxies and measure their redshifts from BOSS spectra, and Lee et al. (2008) did the same to measure stellar properties such as radial velocity, effective temperature, and metallicity from SEGUE spectra.

In this paper, we present an objective and automated method for the detection of line emission in spectroscopic data. This method is designed for the detection of single emission lines (e.g., arising from Lyα emission) within a certain optical spectral range; in our application, 4000–5000 Å. Since the continuum is usually undetected in the observed spectra of most Lyα emitting galaxies, our method mainly focuses on discerning real astronomical signal from coincident noise features by providing quantitative measurements of detection significance. Although the main application of our method is to search for faint emission lines on a zero-mean (continuum subtracted) background, we can extend our method to more generic spectra including radio and infrared spectroscopy, if the continuum can be properly removed.

The outline of our paper is as follows. In § 2, we describe the basic algorithm and present results from Monte Carlo (MC) simulations. In § 3, we apply our technique to 1574 sample spectra obtained using the Hectospec instrument on the MMT. We summarize our findings in § 4.

2. METHOD DESCRIPTION

The problem we address in this paper is how to automatically extract faint emission features (called “detections”) and quantify their significance. For this purpose, we employ commonly used indicators of detection significance: “reliability” and “completeness.” “Reliability” is the probability that a given detection is real (i.e., not resulting from noise). Completeness is the probability that a source of a given flux is detected, given the noise model and the detection method. A completeness equal to 1.0 implies that the feature is detected consistently in all observations of similar depth. In what follows, we describe the mathematical framework used in this paper and then calculate the
reliability and completeness using mock observations from Monte Carlo (MC) simulations.

2.1. Mathematical Formulation

In this section, we present our mathematical framework and definitions of completeness and reliability.

2.1.1. Mathematical Definitions

A signal detection algorithm must quantitatively describe three parts: (1) the intrinsic signals, (2) the measured features, and (3) the detection criteria.

We can parametrize signals using a set of mathematical quantities. We call the set of parameters defining the signal as the “source vector,” \( \tilde{s} \), which spans a “signal space.” Signals are represented as points in this signal space. For example, Gaussian signals can be parametrized by three numbers, i.e., the centroid, width, and height.

We next define the quantities that characterize the detection of signals in observations. In our method, we use a threshold line on spectra (e.g., a flux or signal-to-noise threshold) and measure the width of features lying above the threshold line. In this case, the height of the threshold, \( \Theta_D \), and the width of features selected by the threshold, \( \Delta_D \), will be used to define a “detection” as shown in Figure 1. We call the set of parameters used to define the detection the “detection vector,” denoted by \( \xi = \{\Theta_D, \Delta_D\} \).

Once source and detection vectors are defined, we can run a test for each signal, \( \tilde{s} \), to determine whether it is “detected” or “undetected.” We denote this test using a “test function,” \( \tau(\tilde{s}; \xi) \), which gives Boolean results:

\[
\tilde{s} = \{s_1, s_2, \ldots\},
\]

\[
\xi = \{\xi_1, \xi_2, \ldots\}, \quad \tau(\tilde{s}; \xi) = \begin{cases} 
1 & \text{if detected,} \\
0 & \text{if not detected.}
\end{cases}
\]

In this mathematical framework, the test function is a mapping from signal space to Boolean values, 0 or 1. The test function has a functional dependence on the detection vector, \( \xi \); i.e., different \( \xi \) can produce different test results. This means that by varying \( \xi \) we can optimize the detection efficiency, reliability, and completeness.

In our analysis, we use a Gaussian profile to represent each source signal. Hence, the source vector can be written as \( \tilde{s} = \{s_1, s_2, \ldots\} \) for \( s(\lambda) = s_1 \exp(-\frac{\lambda^2 - \sigma^2}{2\sigma^2}) \), where \( s(\lambda) \) is a source profile.

Now we take into account the “noise” contribution in our mathematical framework. A given detection vector produces one test result on a given single signal. To study the effects of noise, we describe the source as an MC ensemble

MC realizations of \( \tilde{s} \equiv \{\sigma_1(\tilde{s}), \sigma_2(\tilde{s}), \ldots, \sigma_N(\tilde{s})\} \), (3)

where \( \sigma_i(\tilde{s}) \) is the \( i \)-th MC realization of \( \tilde{s} \). We can adopt various noise models to produce MC ensembles for \( \tilde{s} \). In our analysis, we choose simple Gaussian noise, assuming that each pixel is independent, and use a search window of 4000–5000 Å. The sky is nearly constant within this wavelength range, with the exception of two relatively weak telluric emission lines, i.e., Hg I 4047 Å and 4358 Å.

Rewriting the test result, \( \tau(\sigma_i(\tilde{s}); \tilde{\xi}) \), of \( \sigma_i(\tilde{s}) \) in a simpler form,

\[
\tau_i(\tilde{s}; \tilde{\xi}) \equiv \tau(\sigma_i(\tilde{s}); \tilde{\xi}),
\]

we can summarize the presented mathematical representations as

1. \( \tilde{\xi} \): detection vector of parametrized detection criteria,
2. \( \tilde{s} \): source vector of parametrized source signal,
3. \( \sigma_i(\tilde{s}) \): an MC realization (simulated line profile) of source vector, \( \tilde{s} \),
4. \( \tau_i(\tilde{s}; \tilde{\xi}) \): a test result from an MC realization, \( \tau(\sigma_i(\tilde{s}); \tilde{\xi}) \).

2.1.2. Completeness and Reliability

Now we derive completeness and reliability from the above formulation. A series of \( N \) MC realizations for \( \tilde{s} \) yields \( N \) test results, \( \{\tau_1, \tau_2, \ldots, \tau_N\} \), for a given \( \tilde{\xi} \). Hence, the detection probability (or completeness) can be defined as the ensemble average of these test results.
Completeness \( D(\tilde{s}; \tilde{\xi}) = \langle \tau_i(\tilde{s}; \tilde{\xi}) \rangle_{\text{MC}} \),
(5)
where the “MC” bracket represents the Monte Carlo average.

To derive reliability, we need to quantify the probability of false detections, i.e., detections resulting from noise. If the source signal is a null, \( \tilde{s} = 0 \), all detections are false and the result of contiguous (and perhaps correlated) noise spikes. Therefore, the false detection probability is the detection rate of a null signal; i.e.,
\[
F(\tilde{\xi}) = D(\tilde{s} = 0; \tilde{\xi}).
\]
(6)

This false detection rate is an intrinsic limitation of any detection criteria. A conservative detection vector can reduce this rate but this also suppresses the detection of weak but real source signals. Detection vectors must be chosen to maximize completeness, while minimizing false detections.

Simply defined, reliability is the complement of the false detection probability, \( 1 - F(\tilde{\xi}) \). However, since \( 1 - F(\tilde{\xi}) \) is constant (i.e., an MC averaged value), this would imply that different S/N (signal-to-noise ratio) detections such as 10σ or 100σ detections in fact have the same reliability value, which would be misleading. Reliability is therefore more appropriately defined as the complement of the false detection probability density.

We assume a source vector, \( \tilde{s} \), its MC ensemble, \( \sigma_i(\tilde{s}) \), and its test results, \( \tau_i(\tilde{s}; \tilde{\xi}) \). According to the test results, we divide each ensemble \( \sigma_i(\tilde{s}) \) into two disjoint “detected” and “undetected” subsets denoted by \( \sigma_i^D(\tilde{s}) \) and \( \sigma_i^U(\tilde{s}) \). These two subsets satisfy
\[
\{ \sigma_i(\tilde{s}) \} = \{ \sigma_i^D(\tilde{s}) \} \cup \{ \sigma_i^U(\tilde{s}) \},
\]
(7)
\[
\emptyset = \{ \sigma_i^D(\tilde{s}) \} \cap \{ \sigma_i^U(\tilde{s}) \},
\]
(8)
\[
D(\tilde{s}; \tilde{\xi}) = n \{ \sigma_i^D(\tilde{s}) \} / n \{ \sigma_i(\tilde{s}) \},
\]
(9)
where we use braces to denote the sets collecting their MC instances and \( n \{ A \} \) represents the number of elements for the set \( A \). Equations (7)–(9) are simply a restatement of the definition of the test function; the completeness of equation (5) is now rewritten as equation (9).

We use the term, \( \sigma_i(\tilde{s}) \), to represent one simulated MC line profile for \( \tilde{s} \). We can estimate the corresponding parameters \( \tilde{s}_i \) for each \( \sigma_i(\tilde{s}) \) by fitting the profile with a Gaussian model. We refer to this estimation process as “reprojection.”

Thus, for a given source vector, \( \tilde{s} \), we have its MC ensemble \( \{ \sigma_i(\tilde{s}) \} \) of simulated profiles. By reprojection, we obtain its reprojected MC ensemble, \( \{ \tilde{s}_i \} \), in signal space. From this distribution of \( \tilde{s}_i \), we can obtain its corresponding probability density function, \( \tilde{f}(\tilde{x}; \tilde{s}) \), for \( \tilde{s} \), where \( \tilde{x} \) is an independent variable for the density function in signal space. All the three processes, MC realization, reprojection, and probability density function generation can be depicted as
\[
\tilde{s} \rightarrow \{ \sigma_i(\tilde{s}) \} \leftrightarrow \{ \tilde{s}_i \} \rightarrow \tilde{f}(\tilde{x}; \tilde{s}).
\]
(10)

We denote the reprojection process by “\( \leftrightarrow \)” due to incompleteness of the line fitting process. Generally, the fitting procedure depends on a root-finding algorithm. There are inevitable fitting anomalies such as converging to a wrong solution or failing to converge within the root finding trials. Therefore, \( \{ \sigma_i(\tilde{s}) \} \) and \( \{ \tilde{s}_i \} \) do not necessarily represent a one-to-one mapping of each other. However, when the fitting process shows fair correspondence between \( \{ \sigma_i(\tilde{s}) \} \) and \( \{ \tilde{s}_i \} \), we can represent \( \{ \sigma_i(\tilde{s}) \} \) by its density function, \( f(\tilde{x}; \tilde{s}) \), as a practical estimate in signal space. In this argument, the detected subset, \( \{ \sigma_i^D(\tilde{s}) \} \), can be described by a probability density function in signal space with an acceptable fitting performance:
\[
\{ \sigma_i^D(\tilde{s}) \} \rightarrow \tilde{f}_D(\tilde{x}; \tilde{s}; \tilde{\xi}).
\]
(11)

We call \( \tilde{f}_D(\tilde{x}; \tilde{s}; \tilde{\xi}) \) the Detection Probability Density Function (DPDF) for \( \tilde{s} \). We note that, by the reprojection process and the definition of DPDF, all observed detections and quantities derived from MC simulations are now plotted in signal space; in our implementation, the parameter space for a Gaussian profile.

From equation (5), the normalization of DPDF for \( \tilde{s} \) is
\[
D(\tilde{s}; \tilde{\xi}) = \int \tilde{f}_D(\tilde{x}; \tilde{s}; \tilde{\xi}) d\tilde{x}.
\]
But, for convenience, we rescale the normalization condition as
\[
1 = \int \tilde{f}_D(\tilde{x}; \tilde{s}; \tilde{\xi}) d\tilde{x},
\]
(12)
where its original normalization is replaced by the conditional probability, \( D(\tilde{s}; \tilde{\xi}) \tilde{f}_D(\tilde{x}; \tilde{s}; \tilde{\xi}) \).

The sourceless DPDF, \( \tilde{f}_D(\tilde{x}; \tilde{s} = 0; \tilde{\xi}) \), is the “false detection probability density” and its normalization is the “false detection probability” in equation (6),
\[
F(\tilde{\xi}) = \int D(\tilde{s} = 0; \tilde{\xi}) \tilde{f}_D(\tilde{x}; \tilde{s} = 0; \tilde{\xi}) d\tilde{x}.
\]
(13)

Therefore, the distribution of \( \tilde{f}_D(\tilde{x}; \tilde{s} = 0; \tilde{\xi}) \) is the zone of false detections in signal space. To derive the final mathematical expression for reliability, we assume an observed spectrum, \( \sigma_o \), and its reprojection, \( \tilde{\sigma}_o \). As a trivial case of \( \tilde{f}_D(\tilde{o}; \tilde{s} = 0; \tilde{\xi}) = 0 \), any observed detections would be deemed fully reliable due to the perfect decoupling of true and false detections. Therefore, at least, we can assign reliability \( = 1 \) to all observed detections of \( \tilde{f}_D(\tilde{o}; \tilde{s} = 0; \tilde{\xi}) \) = 0.

In the case where \( \tilde{f}_D(\tilde{o}; \tilde{s} = 0; \tilde{\xi}) \neq 0 \), we can have a couple of definitions for reliability. One of the simplest definitions would be to reject all of those partially reliable detections,
\[
\text{Reliability} = \begin{cases} 1 & \text{if } \tilde{f}_D(\tilde{o}; \tilde{s} = 0; \tilde{\xi}) = 0, \\ 0 & \text{if } \tilde{f}_D(\tilde{o}; \tilde{s} = 0; \tilde{\xi}) \neq 0. \end{cases}
\]

This definition can be powerful when focusing on detecting strong emission lines. However, it is clear that we need a better definition if we hope to detect weaker emission lines as well.
The quick-and-dirty extension can be to use the profile shape of \(\tilde{f}_D(\tilde{x}; \tilde{s} = 0, \xi)\),

\[
\text{Reliability} = 1 - \tilde{f}_D(\tilde{o}; \tilde{s} = 0, \xi)/f_{\text{max}},
\]

where \(f_{\text{max}}\) is the maximum value of \(\tilde{f}_D(\tilde{x}; \tilde{s} = 0, \xi)\). In this definition, we can assign fractional reliability to the detections of \(f_D(\tilde{o}; \tilde{s} = 0, \xi) \neq 0\). Expanding this to the full probability density function in order to obtain the correct reliability fraction, we choose a “contour” in the signal space to separate “reliable” from “unreliable” detections. For instance, we define this contour to be \(C(\mu)\) such that:

\[
\mu \equiv \int_{C(\mu) \geq f_D(\tilde{x}; \tilde{s} = 0, \xi)} \tilde{f}_D(\tilde{x}; \tilde{s} = 0, \xi)d\tilde{x},
\]

\[
1 - \mu = \int_{f_{\text{max}} \geq f_D(\tilde{x}; \tilde{s} = 0, \xi)} \tilde{f}_D(\tilde{x}; \tilde{s} = 0, \xi)d\tilde{x},
\]

where we have the boundary conditions of \(C(0) = 0\) and \(C(1) = f_{\text{max}}\). If we choose \(\mu = 0.99\), then any observed detection falling outside this region has a less than 1% chance of being a false detection; its “reliability” is therefore \(\geq 99\%\). Therefore, we can define the reliability at \(\tilde{x}\) as,

\[
\text{Reliability} = 1 - \tilde{F}_D(\tilde{x}; \tilde{s} = 0, \xi).
\]

\[
\tilde{F}_D(\tilde{x}; \tilde{s} = 0, \xi) \equiv \int_{f_D(\tilde{x}; \tilde{s} = 0, \xi) \geq C(\mu)} \tilde{f}_D(\tilde{x}; \tilde{s} = 0, \xi)d\tilde{x}.
\]

In practice, we do not have to calculate \(\tilde{F}_D(\tilde{x}; \tilde{s} = 0, \xi)\) for all \(\tilde{x}\) in signal space. We only need a reliability threshold for the adopted detection vector and its corresponding contour line such as \(C(0.05)\) or \(C(0.01)\) for 95% or 99% reliability.

Before we move on to the implementation, we discuss two points related to our mathematical framework. The first is about another possible variant of reliability,

\[
\text{Reliability} = 1 - \tilde{F}_D(\tilde{x}; \tilde{s} = 0, \xi).
\]

Guaranteed the failure rate of a detection is very low, such as \(\tilde{F}(\xi) = 0.01\). If the rejection rate of noise signals is very high, then the reliability of a detection may be underestimated by the definition in equation (17). A weighted approach (provided by eq. (19)) may have some advantages in this case. However, we choose to ignore any detection that is indistinguishable from noise, and hence conservatively adopt the definition in equation (17) for the present study.

The second is about a more general interpretation of the DPDF. Since we have focused on the test of detection, we have only needed the sourceless DPDF and derived equation (18). However, we can derive the DPDF, \(\tilde{F}_D(\tilde{x}; \tilde{s} = \tilde{s}_1, \xi)\), for a general source vector, \(\tilde{s}_1\). For an observed detection, \(\tilde{o}\), we can calculate \(\tilde{F}_D(\tilde{x} = \tilde{o}; \tilde{s} = \tilde{s}_1, \xi)\) as well as \(\tilde{F}_D(\tilde{x} = \tilde{o}; \tilde{s} = 0, \xi)\). As explained before, the latter, \(\tilde{F}_D(\tilde{x} = \tilde{o}; \tilde{s} = 0, \xi)\), provides the reliability of the detection. Similarly, but more interestingly, the former, \(\tilde{F}_D(\tilde{x} = \tilde{o}; \tilde{s} = \tilde{s}_1, \xi)\), provides the probability that the observed feature, \(\tilde{o}\), is originated from the underlying source, \(\tilde{s}_1\). Therefore, the on-source DPDFs provide all the possible candidates of the underlying sources for a given observed detection with their own probabilities, though we do not have to use this extra information.

By following the basic descriptions above, the automated pipelines we will implement can extract detections from spectra and assign the completeness and reliability values to each detection. By restricting the reliability or completeness, we can finalize our detections. This is the basic frame for our automated detection method. The following sections will present the specific steps to build the detection pipelines.

2.2. Implementation

2.2.1. Standardization and Generation of the Pseudospectrum

Generally astronomical spectra are imaged on Charge-coupled Device (CCD) cameras. One-dimensional spectra are extracted from the two-dimensional images and typically represented by three data arrays of wavelength, spectrum profile, and inverse variance with pixel indices. Their pixel sampling rate, S/N, and flux calibration vary according to weather conditions, instruments, and optics. Therefore, we need steps to standardize the observation-dependent spectra for use with a generic detection algorithm.

In our detection method, we need two requirements of standardization: (1) uniform wavelength sampling and (2) normalization of variance to unity. The first is necessary to interchange between wavelength scale and pixel scale. The pixel scale is an easier and more standard unit to work on data-processing problems, while scientific requirements and constraints are written in terms of wavelength. The conversion between the two scales is consistent and trivial when the sampling rate is uniform. Therefore, for convenience and consistency, uniform wavelength sampling is required in our method. In practice, if pixel sampling rates vary within \(\pm 10\%\) in a target spectral range, the uniform resampling is not necessary.

The second is to normalize the noise level to unity. Since the noise level is a complex result of observational effects (e.g., due to instruments, optics, detectors, weather, etc.), we cannot model all possible situations in our MC simulations. Therefore, we normalize the observed spectra by their standard deviations to equalize the noise to unity. That is, if we denote the original data set of wavelength, spectrum, standard deviation of noise by \((\lambda(x), s(x), \sigma(x))\), then the normalized data set will be \((\lambda(x), s(x)/\sigma(x), 1)\). We call this normalized spectrum, \(s(x)/\sigma(x)\), the “pseudospectrum.” Our Monte Carlo analysis is built for these pseudospectra. The pseudospectrum is not in physical units (e.g., flux density), but is effectively a dimensionless...
signal-to-noise ratio. The profile shape in the pseudospectrum is also deformed during the normalization process. This deformation is relatively small if the noises in neighboring pixels are similar. In practice, the shapes of features in the pseudospectra are similar to those in the original spectra as long as the noise varies smoothly under them, as is the case in spectral regions that do not include strong telluric lines.

To describe the meaning of a pseudospectrum, we start by summarizing the quantities in mathematical terms. In the spectral range of \( (\lambda_0, \lambda_0 + \Lambda) \), the flux, pseudoflux (integration of the pseudospectrum), and S/N can be written as:

\[
\text{Flux} = \int_{\lambda_0}^{\lambda_0 + \Lambda} s(\lambda) d\lambda = \sum_{x=0}^{N} s(x),
\]

\[
\text{Pseudoflux} = \int_{\lambda_0}^{\lambda_0 + \Lambda} \frac{s(\lambda)}{\sigma(\lambda)} d\lambda = \sum_{x=0}^{N} \frac{s(x)}{\sigma(x)},
\]

\[
\frac{S}{N} = \frac{\int_{\lambda_0}^{\lambda_0 + \Lambda} s(\lambda) d\lambda}{\left(\int_{\lambda_0}^{\lambda_0 + \Lambda} \sigma^2(\lambda) d\lambda\right)^{1/2}} = \frac{\sum_{x=0}^{N} s(x)}{\left(\sum_{x=0}^{N} \sigma^2(x)\right)^{1/2}} = \frac{\sum_{x=0}^{N} s(x)}{\sqrt{N} \sigma_{\text{rms}}},
\]

where \( \lambda(x = 0) = \lambda_0 \), \( \lambda(x = N) = \lambda_0 + \Lambda \), and \( \sigma_{\text{rms}} \) is a root-mean-square of \( \sigma(x) \); generally, \( \Lambda \) is a couple of times the FWHM of a typical emission line. For the trivial case of uniform noise, \( \sigma(x) = \sigma_0 \), the mean and rms values are the same, \( \sigma(x) = \bar{\sigma} = \sigma_{\text{rms}} = \sigma_0 \), where \( \bar{\sigma} \) is the mean of \( \sigma(x) \).

**Fig. 2.—**The left-top panel shows a Gaussian random noise spectrum with \( \sigma = 1 \) (gray) and a Gaussian source signal with pseudoflux = 37.6 and FWHM = 7.06 (black; or \( f_{\text{max}} = 5 \) and \( \sigma = 3 \)). The right-top panel shows the combined profile of the source and noise. The y-axis is in the pseudospectrum scale and the x-axis in the pixel scale. With the given threshold lines, \( \Theta_D = 1.5 \) (middle-left) and 2 (bottom-left), we can make the detection mask assigning “1” where the signal is larger than threshold (middle- and bottom-right panels). This detection mask results in detection blocks of different widths \( \delta \). For each random realization, we measure detected widths and stack them to the probability function, \( p_\lambda(\delta; \Theta_D) \), at each pixel position. For a given position, \( \lambda \), and a detection threshold, \( \Theta_D \), the sum of all stacked counts for each detected \( \delta \) is exactly the same as the total number of random realizations; i.e., \( \sum_{\delta=0}^{\infty} p_\lambda(\delta; \Theta_D) = 1 \). The detected width at \( \lambda = 200 \) is larger than at other sourceless positions. This property is the key feature for our detection method.
If we extend this trivial case to “quasi-uniform” cases of \( \sigma(x) \sim \bar{\sigma} \sim \sigma_{\text{rms}} \), the pseudoflux can be approximated as

\[
P \text{Pseudoflux} \sim \sum_{x=0}^{N} s(x) / \bar{\sigma} \sim \sqrt{N} \times \text{S/N}. \tag{20}
\]

Unless our target line emission falls in close proximity to a bright sky line, equation (20) generally holds. The pseudospectrum is (roughly) a scaled measure of the S/N per pixel.

### 2.2.2. Threshold, Detection Mask, and Detected Width

We now choose a detection threshold and create a Boolean detection mask from the pseudospectrum, where “1” corresponds to pixel values falling above the threshold, and “0” corresponds to values below it. Figure 2 shows two threshold lines, \( \Theta_D = 1.5 \) (middle panels) and 2.0 (bottom panels) and their detection masks, where \( \lambda \) is on a uniformly resampled pixel scale as explained in the previous section. The top-left panel shows a spectrum composed of Gaussian random noise with \( \sigma = 1 \) (gray line) and a Gaussian emission feature of width 3 and height 5 centered at pixel \( \lambda = 200 \) (black line). The top-right panel shows their combined signal as a pseudospectrum. Using the detection mask, we identify clustered blocks of detected pixels and measure their sizes (widths). We call this block size the “detected width” \( \delta \).

For a given source vector \( \vec{s} \), we obtain one detected width value from each random noise realization. From a series of MC realizations, we obtain an MC ensemble of the detected widths denoted by \( \{ \delta_i(\vec{s}, \Theta_D) \} \). These measured detected widths are stacked and reduced to probability function:

\[
\{ \delta_i(\vec{s}, \Theta_D) \} \rightarrow p(\delta; \vec{s}, \Theta_D) \tag{21}
\]

with the normalization of \( \sum_{\delta} p(\delta; \vec{s}, \Theta_D) = 1 \). We call this probability function as Detected Width Probability Function (DWPF).

The two ensembles of on-source \( \{ \delta_i(\vec{s}, \Theta_D) \} \) and off-source \( \{ \delta_i(\vec{0}, \Theta_D) \} \) (i.e., on and off the emission line feature) width measurements are the fundamental data sets used to define the reliability and completeness in our detection method. From the ensembles, we obtain the corresponding DWPFs, \( p(\delta; \vec{s}, \Theta_D) \) and \( p(\delta; \vec{0}, \Theta_D) \). Figure 3 shows the two DWPFs, \( p_{\lambda=200}(\delta; \Theta_D) \) and \( p_{\lambda=300}(\delta; \Theta_D) \), measured at the two different positions, \( \lambda = 200 \) and 300, for the given source vector \( \vec{s} \) shown in Figure 2. By definition, the DWPF measured at \( \lambda = 200 \) is the on-source DWPF for \( \vec{s} \),

\[
p(\delta; \vec{s}, \Theta_D) = p_{\lambda=200}(\delta; \Theta_D). \tag{22}
\]

We can obtain the off-source DWPF by rerunning the MC simulations for \( \vec{s} = \vec{0} \). However, since there is no contribution from the source signal at \( \lambda = 300 \) (or sufficiently away from the signal), we can measure the off-source DWPF simultaneously as

\[
p(\delta; \vec{0}, \Theta_D) = p_{\lambda=300}(\delta; \Theta_D). \tag{23}
\]

The clear difference between \( p_{\lambda=200}(\delta; \Theta_D) \) and \( p_{\lambda=300}(\delta; \Theta_D) \) shown in Figure 3 is the key feature to set our detection criteria.

### 2.3. Completeness

From equations (22) and (23), we can directly derive completeness from the definition of “detection.” Figure 3 shows that the probability of \( \delta \geq 4 \) for the off-source ensemble is quite small, while most of the on-source ensemble have \( \delta \geq 4 \). From this key observation, we choose a detection width threshold, \( \Delta_D \), and define a “detection” as a case where the measured width is larger than this threshold, \( \delta \geq \Delta_D \). If we parametrize a Gaussian profile using its pseudoflux and full-width-half-maximum (FWHM), the source vector and detection vector of our method can be denoted by

\[
\vec{s} = \{ \text{Pseudoflux, FWHM} \}, \tag{24}
\]

\[
\vec{\xi} = \{ \Theta_D, \Delta_D \}. \tag{25}
\]

We now write the detection probability and false detection probability defined in equations (5) and (6) as

\[
D(\vec{s}; \Theta_D, \Delta_D) = \sum_{\delta \geq \Delta_D} p(\delta; \vec{s}, \Theta_D), \tag{26}
\]

\[
F(\Theta_D, \Delta_D) = \sum_{\delta \geq \Delta_D} p(\delta; \vec{0}, \Theta_D). \tag{27}
\]
Using equations (22) and (23), Figure 3 shows the detection and false detection probabilities, $D(s; \Theta_D = 1.5, \Delta_D = 4) = 0.99965$ and $F(\Theta_D = 1.5, \Delta_D = 4) = 0.00005$, where the source vector, $s = \{38, 7.1\}$, is given on the top-left panel in Figure 2. Since $D(s; \Theta_D, \Delta_D)$ is the completeness for $s$, we finally have

$$
\text{Completeness} \equiv \sum_{\delta \in \Delta_D} p(\delta; s, \Theta_D).
$$

For each set of detection criteria $\{\Theta_D, \Delta_D\}$, we can calculate the completeness values for all source vectors $s = \{\text{Pseudoflux, FWHM}\}$.

Figure 4 shows the completeness contours in signal space for various $\{\Theta_D, \Delta_D\}$, derived from $10^5$ MC realizations. The blue dotted and dashed lines show 0.95 and 0.99 completeness contours. Strong signals and high pseudofluxes guarantee high completeness as we would expect. For the same pseudoflux, the profiles with smaller FWHMs are more likely to be detected, since their peaks are higher. The limit of the narrowest width is enforced by the threshold width, $\Delta_D$. All of these properties form the bow-shaped contours of completeness in the signal space.

As $\Theta_D$ and $\Delta_D$ increase, the 0.95 and 0.99 contour lines move to the right side reducing the coverage area, but ensuring smaller false detection probabilities. This tradeoff needs to be considered when choosing the final detection vector. The data we will analyze in § 3 from the MMT/Hectospec have the spectral resolution of FWHM = 6 Å and pixel sampling rate of 1.1 Å/pixel. Based on Figure 4, we choose $\{\Theta_D, \Delta_D\} = \{1.5, 4\}$ as the final detection vector. This corresponds to the detection of the core of an emission line with 4 pixels lying at $\geq 1.5\sigma$, or roughly a total core flux of $\geq 3\sigma$.

### 2.4. Reliability

As described in § 2.1, we need to investigate each MC realization to derive reliability. Since we have one detection width for one MC realization, we can use this one-to-one correspondence to represent each MC realization,

$$
\bar{s} \rightarrow \sigma_i(\bar{s}) \sim \{\delta_i(s, \Theta_D)\}.
$$

---

**Fig. 4.**—The completeness (blue contours) and false detection probabilities (green) from $10^5$ random noise realizations for various detection parameters; $(\Theta_D, \Delta_D)$. The y-axis of FWHMs is in the pixel scale. By adopting more conservative detection criteria, i.e., higher $\Theta_D$ or higher $\Delta_D$, we can reduce the false detection probability but, at the same time, we lose more completeness; i.e., losing more chances to detect faint source signals. We need to consider this tradeoff when choosing detection parameters. To apply our Monte Carlo results to the MMT/Hectospec data presented in § 3, we choose $\Theta_D = 1.5$ and $\Delta_D = 4$ shown in Fig. 5. See the electronic edition of the PASP for a color version of this figure.
Then, the detected and undetected subsets can be represented by the two \( \delta \geq \Delta_D \) and \( \delta < \Delta_D \) as

\[
\{ \sigma^D_i(\delta) \} \sim \{ \delta_{\geq \Delta_D}(\delta, \Theta_D) \}
\]

(30)

\[
\{ \sigma_i(\delta) \} \sim \{ \delta_{< \Delta_D}(\delta, \Theta_D) \}
\]

(31)

Finally, during reprojection, a Gaussian is fit to features with \( \delta \geq \Delta_D \).

\[
\{ \delta_{\geq \Delta_D}(\delta, \Theta_D) \} \rightarrow \{ \hat{\delta}_{\geq \Delta_D} \}
\]

(32)

The top panel of Figure 5 shows three reprojection samples for

1. C: the signal of pseudoflux = 70 and FWHM = 8,
2. R1: false detections for \( \Theta_D = 1.5 \) and \( \Delta_D = 4 \),
3. R2: false detections adding +1\( \sigma \) to baseline; its effective threshold \( \Theta_D = 0.5 \),

with the completeness contours for \( \Theta_D = 1.5 \) and \( \Delta_D = 4 \).

The completeness of the sample C (red points) is equal to 1. The most probable point (black cross) of C coincides with its original source signal of pseudoflux = 70 and FWHM = 8. This is a typical reprojection pattern for a strong signal. The R1 sample shows the reprojections of false detections, i.e., zero source signal, for \( \Theta_D = 1.5 \) and \( \Delta_D = 4 \). This is the theoretical limit of false detections. Practically, however, baseline estimation is rarely perfect and we may need to allow for some additional systematic uncertainty. Generally both sky subtraction and continuum estimation affect this baseline uncertainty. Underestimating the baseline (i.e., undersubtracting the sky background or continuum from spectra) will artificially elevate the noise above the threshold and result in an excess number of false detections. Conversely, oversubtracting the background will cause real features to be missed, though the false detection rate may be underestimated. In order to estimate the impact of errors in baseline subtraction, we added a systematic 1\( \sigma \) offset to the baseline and measured the resulting detections and their widths. We refer to this sample, which has an effective detection threshold of \( \Theta_D = 0.5 \), as the “R2 sample.” The R2 sample results in a false detection probability \#300 times larger than that of the R1 sample. Since the observed false detections (§ 3.2.1) are much lower than the false detection counts of R2 sample, we can use the R2 sample as a generous upper limit to the false detection rate (i.e., a conservative limit of real detections) to cover the conditions of poor sky subtraction or continuum subtraction which increase the uncertainty of baseline estimates.

The bottom panel of Figure 5 shows the final reliability contours defined in equations (18) and (17) for the R2 sample. This panel is our key result, and is used to assign reliability and completeness to each detected feature. Since R2 is an upper limit, the R2 contours provide more conservative reliability (i.e., more generous false detections probability density). The actual distribution of false detections likely lies between R1 and R2.

2.5. Incomplete Reprojections: Fitting Anomalies

As previously mentioned, most root finding or optimization routines suffer at times from converging to local minimum (finding a wrong solution) or never converging to any solution (failing to find solution). The reprojection process, therefore, is not complete due to these inevitable fitting anomalies. Here, we investigate these anomalies to quantify the reprojection process.

In our method, we use MPFIT packages (Markwardt 2009) to fit a Gaussian line profile to each detected feature in order to estimate the source parameters. For a given detected width, \( \delta \),
we add a width, W, on both of the left and right sides of the detection to define a fitting range of \( \delta + 2W \). When we run MPFIT with default settings and \( W = 3 \), we find three kinds of fit anomalies: (1) E1: centroid is out of the fitting range, (2) E2: negative flux, and (3) E3: width is narrower than 1 pixel (which is not possible for real features given the oversampled data). Figure 6 shows these fitting results for the R1 and R2 samples. Instead of pseudoflux and FWHM, we use the height, \( F_{\max} \), and Gaussian sigma width, \( \sigma \), to show the anomalies more clearly. They are related as pseudoflux \( = 2.5066 \sigma F_{\max} \) and FWHM \( = 2.3548 \sigma \). Since we search for emission lines within the detected width, none of the three anomalies are valid fits.

In order to minimize fitting anomalies, we investigated the impact of various constraints to the postdetection line-fitting algorithm and converged on the following constraints: (1) positive flux only, (2) \( W = 3 \), and (3) the centroid must be located in the fit range; applied in the fitting routines. Based on these constraints, we obtained near-complete reprojections; zero anomalies for the R1 sample and four E3 anomalies for the R2 sample from \( 10^6 \) MC realizations.

### 2.6. Number Statistics of False Detections

In this section, we discuss how the false detection statistics depend upon the size of the spectral range over which the detection search algorithm is run. We assume that \( \Lambda \) is a spectral range of interest and \( p_N(\delta; \Theta_D) \) is a noise DWPF as described by equation (23). Since the source signal term is zero, \( p_N(\delta; \Theta_D) \) only depends on \( \Theta_D \). Figure 7 shows \( p_N(\delta; \Theta_D) \) versus \( \Theta_D \). This \( p_N(\delta; \Theta_D) \) is the probability of a detection of width \( \delta \) at a certain sourceless position. Therefore, over the range \( \Lambda \), we can conjecture that the possible false detection counts are the statistic of binomial trials using the probability of \( p_N(\delta; \Theta_D) \). This is true if we take into account the redundant counts of \( \delta \) by diving \( p_N(\delta) \) by \( \delta \). Therefore, the number of false detections in the spectral range, \( \Lambda \), can be expressed as:

\[
N_{\text{false}}(\delta; \Theta_D, \Lambda) \approx B\left( \Lambda, \frac{p_N(\delta; \Theta_D)}{\delta} \right)
\]

\[
N_{\text{false}}(\Delta_D, \Theta_D, \Lambda) \approx B\left( \Lambda, \sum_{\delta \leq \Delta_D} \frac{p_N(\delta; \Theta_D)}{\delta} \right)
\]

where \( B(N, p) \) represents a binomial distribution of \( N \) trials with a probability \( p \).

To verify our argument, we produced \( 10^5 \) MC realizations with \( \Lambda = 500 \) and \( 2000 \) and measured the number of noise detections and compared these with their binomial distributions. Figure 8 shows the false detection counts from MC realizations (diamonds) and the binomial distributions (lines) from equation (33) for \( \Theta_D = 0.5 \) and \( \Delta_D = 4 \). We used \( \Theta_D = 0.5 \) to obtain enough counts for comparison, since \( p_N(\delta; \Theta_D = 1.5) \) is too small. This good agreement is not a surprising result because \( p_N(\delta; \Theta_D) \) itself is derived from the same setups of the MC simulation. This binomial description for the number statistics of false detections is in a sense rephrasing the definition of the DWPF. We note that, if \( \Lambda \approx \delta \), the binomial description fails because within the length of \( \delta \) the probability is not independent. In general, however, the search window is much larger than the width of an emission line, \( \Lambda \gg \delta \). Therefore, for most cases, we can easily estimate the false detection counts from the DWPF, without running additional MC simulations, we can easily estimate the false detection counts from DWPF.

For example, as we show in § 3, when we search for LAEs in the range from 4000 to 5000 Å with 1 Å/pixel sampling, i.e., \( \Lambda = 1000 \), with the detection criteria, \( \Theta_D = 1.5 \) and \( \Delta_D = 4 \), the false detection probability, \( \sum_{\delta \leq \Delta_D} \frac{p_N(\delta; \Theta_D)}{\delta} \), is \( 1.78 \times 10^{-5} \). Hence, the average false detection counts for each spectrum are \( 0.0178 \pm 0.1334 \). Since the false detection counts are much lower than 1 for a single spectrum, this detection threshold is sufficient to suppress most false detections. Since our total
A description of the HSRED also can be found in Kochanek et al. (2012). The central wavelength of IA445 is 4458 Å and its FWHM width 201 Å. The full details about the reductions of the IA445 image are presented in Prescott et al. (2008) and A. Dey et al. 2015, in preparation. Briefly, the total exposure time was 3 hr. The images were reduced using the SDFRED software (Yagi et al. 2002; Ouchi et al. 2004). We registered and resampled the IA445 image to match the B and R imaging from NDWFS. We retrieved ≈38,600 sources from the IA445 image using SExtractor with the 5σ limit of 26.5 AB mag (Bertin & Arnouts 1996).

3.1.2. MMT/Hectospec Observations

We observed selected candidates using MMT/Hectospec to confirm the presence of Lyα and obtain spectroscopic redshifts of LAEs. Hectospec is a 300-fiber and 1° field-of-view multi-object spectrograph at the MMT Observatory (Fabricant et al. 2005). We used the program XFITFIBS to assign optical fibers to science and calibration targets in our field and observed 7 configurations. We used the 270 line/mm grating blazed at 5200 Å with a resolution of FWHM ≈ 6.2 Å.

We observed 1574 LAE candidates using seven configurations over six nights as summarized in Table 1. The seeing was >1.6 in windy weather on 2012 May 16 and 17. The effective exposure times for M2 and J1 configurations are, therefore, smaller than the presented ones and their data qualities are poor. On the other days, the weather conditions were acceptable and the seeing was 0.64–1.0″. The details about the MMT/Hectospec observations and reductions will be presented by S. Hong et al. (2015; in preparation). Briefly, we use the HSRED package, a modified version of the Sloan Digital Sky Survey (SDSS) pipeline, written in Interactive Data Language (IDL) by Richard Cool.4 A description of the HSRED also can be found in Kochanek et al. (2012). The basic reduction of subtracting bias, flat-fielding, calibrating arc, and sky subtraction are done by the HSRED. We apply our detection method to the output spectra from the HSRED pipelines.

3.2. Results

We split the application of our automated pipeline into two phases: (1) raw detections and (2) a customized selection from the raw detections. During the first phase, the pipeline assigns all detected features their own reliability and completeness values. In the second phase, we set a specific reliability threshold to identify “reliable” detections and maximize the exclusion of false positives.

4 Please see http://www.mmto.org/rcool/hsred/ for more information.
3.2.1. Raw Detections

For illustration purposes, we start by focusing on a single pointing—the PM1 configuration. We observed 244 LAE candidates, of which 179 have raw detections in the search wavelength range 4000–5000 Å for the criteria of $\Theta_D = 1.5$ and $\Delta_D = 4$. The number of false detections for R1 (in § 2.4), representing the case of perfect baseline estimation, is $0.016 \pm 0.13$; for all 244 spectra, $3.9 \pm 2.0$. For R2 representing the case of poor baseline estimation, we expect $5.6 \pm 2.4$ false detections per spectrum, for all 244 spectra, this corresponds to $1370 \pm 37$. However, since the sample R2 assumes that the baseline is overestimated by $+1\sigma$ through the whole

![Graphs and tables showing raw detections and statistics.](image)

**Fig. 9.**—Raw detections resulting from our analysis of the data from MMT/Hectospec configuration PM1. The top panels show the observed source vectors (Gaussian fits of the detections) for the first, second, and third detections in each spectrum, as defined in the text, with the reliability and completeness contours shown, as in Fig. 5. The bottom panels show the histograms of the measured centroids with the green lines representing the throughput of the IA445 filter. The red dotted lines are the sky emission lines, Hg 1 4047 Å and 4358 Å. See the electronic edition of the PASP for a color version of this figure.

### TABLE 1

| Configuration (label) | R.A.$^a$ (J2000.0) | Decl.$^a$ (J2000.0) | Date$^b$ | Exposure time (s) | No. of LAE targets$^c$ |
|-----------------------|---------------------|---------------------|----------|-------------------|------------------------|
| PM1                   | 14 32 23.6          | +33 24 44           | 2009 May 28 | 6 × 1800          | 244                    |
| PM2                   | 14 33 35.9          | +33 27 16           | 2009 May 29 | 6 × 1800          | 238                    |
| M1                    | 14 32 19.7          | +33 24 05           | 2012 May 17 | 6 × 1560          | 226                    |
| M2                    | 14 32 59.8          | +33 23 13           | 2012 May 16 | 6 × 1600          | 221                    |
| M3                    | 14 32 43.1          | +33 22 25           | 2012 May 21 | 3 × 1800          | 232                    |
| J1                    | 14 32 24.9          | +33 25 31           | 2012 Jun 17 | 4 × 1800          | 217                    |
| J2                    | 14 33 36.2          | +33 23 24           | 2012 Jun 21 | 3 × 1800          | 203                    |

*Note.—Units of right ascension are hours, minutes, and seconds, and units of declination are degrees, arcminutes, and arcseconds.

$^a$ The pointing position for each configuration. The field of view is 1° in diameter centered on this position.

$^b$ The seeing was $>1.6''$ in windy weather on 2012 May 16 and June 17. The effective exposure times for M2 and J1 are, therefore, smaller than the presented ones. On the other days, the weather conditions were acceptable and the seeing 0.64–1.0''

$^c$ The number of fibers assigned to the LAE candidates out of the total 300 fibers. The total is 1581 with 7 duplicated targets.
searching spectral range, the estimate from R2 is an exaggerated upper-bound. When the real baseline errors are suppressed or enhanced through the spectral range larger than \(\pm 1\sigma\), the effective spectral length enhancing the \(+1\sigma\) baseline error should be some fraction of the total spectral range. Therefore, accounting for this effective fraction of the enhanced fluctuation, the practical estimates of false detections may be around 0–3 false detections on each spectrum.

The top panels in Figure 9 show the reprojections (Gaussian fits) of the raw detections with the reliability (cyan) and completeness (blue) contours taken from Figure 5. These top panels are the key outputs from our detection method. The bottom panels show the histograms of line centroids alongside the transmission curve of the IA445 filter (green lines). From the false detection statistics presented above, we generally expect a couple of detections on each spectrum. When there is more than one detection in a given spectrum, we use the term “second detections” for secondary features and “third detections” for tertiary features, ordered by pseudoflux. The red points and histogram represent the first detections (first column), the grey the second detections (second column), and the black the third detections (third column). The red dotted lines on the bottom panels represent the sky emission lines, Hg I 4047 Å and 4358 Å.

We find two important results from Figure 9. First, the centroid histogram of the first detections (bottom-left panel) shows a good correlation with the IA445 filter transmission curve, while the other histograms of the second and third detections (bottom-middle and -right panels) do not. This implies that the target selection and our automated method work properly, suggesting that the first detections are likely to be real emission features.

\[
\begin{align*}
\text{D100 (1st detection)} & : \\
\text{D110 (1st detection)} & : \\
\text{D111 (1st detection)} & : \\
\end{align*}
\]

\[
\begin{align*}
\text{D100 (2nd detection)} & : \\
\text{D110 (2nd detection)} & : \\
\text{D111 (2nd detection)} & : \\
\end{align*}
\]

\[
\begin{align*}
\text{D100 (3rd detection)} & : \\
\text{D110 (3rd detection)} & : \\
\text{D111 (3rd detection)} & : \\
\end{align*}
\]

Fig. 10.—Histograms of the detected centroids for the three detection categories, D100, D110, and D111, in PM1. The sum of each column results in the histograms in Fig. 9. Since the expected number of false detections is \(3.9 \pm 2.0\) for perfect sky and continuum subtraction, the 46 and 17 detections in D110 and D111 are likely the result of erroneous baseline fluctuations caused by the practical limitations of sky subtraction. See the electronic edition of the PASP for a color version of this figure.
line detections. Conversely, the poor match for the second and third detections implies that they are more likely to be noise detections. The fact that their reprojections shown in the top-middle and right panels fall within the low reliability contours also supports this argument. A slight excess in the second detections near 4500 Å implies that some of the second detections could be real Lyα emission lines. There are very few real detections among the third detections. Hence, we exclude

![Detection Histogram](image1)

![Spatial Distribution](image2)

![Pseudoflux Spectrum](image3)

![Original Profile](image4)

**Fig. 11.**—The black crosses in the top left panel shows the wavelength locations of all the lines detected for the entire set of LAE candidates; the location of the two emission line detections for LAEA5569 are represented by the blue asterisk. The vertical and horizontal lines represent the locations of the two telluric emission lines within the search wavelength range (Hg I 4047 Å and 4358 Å); both detections for LAEA5569 fall on the locations of these sky lines. The right panel shows the wavelength histogram for all detections in the sample. The bulk of detections lie within the boundaries of the narrow-band filter used to select LAE candidates (green line). The vertical dot-dashed line shows the location of the first detection for LAEA5569, which clearly lies outside the filter selection range. The middle row shows the distribution of LAEs (purple crosses) in a FWHM vs. Pseudoflux plane (left panel) and on the sky (right panel), with the LAEA5569 detection marked as a blue asterisk. The bottom left panel shows the pseudoflux spectrum in the vicinity of the detected feature; it clearly resembles a sky-line residual. The bottom right panel shows the spectrum of LAEA5569 over the full wavelength search range with the emission line detection highlighted and the location of the Hg I sky lines shown as dashed vertical lines. This is a typical example of a case that must be removed by appropriate rejection. See the electronic edition of the *PASP* for a color version of this figure.
all the third detections from the real emission candidates. The second result is that there are many junk detections caused by residuals associated with the improper subtraction of the Hg I 4047 Å and 4358 Å telluric emission lines in all of the first, second, and third detections. The second phase of the pipeline analysis is needed to deal with these junk (sky residual) detections.

To investigate the raw detections further, we categorize the 179 raw detections into three kinds, called D100, D110, and D111. D100 represents the spectra having only first detections, D110 having first and second detections, and D111 having three or more detections. Figure 10 shows the histograms of the three categories, D100, D110, and D111. The sum of each column results in the histograms of Figure 9. The numbers of D110 and D111 detections are 46 and 17. If we assume there is one real emission line in D110 and D111, we have $46 + 2 \times 17$ false detections. When comparing these 80 false positives with the numbers predicted from the R1 and R2 reliability contours ($4/C_2^6$ and $1370/C_3^7$ respectively), we find that the R2

![Detection Histogram]

![Spatial Distribution]

![Pseudoprofile vs. Profile](Laev21116spec)

**Fig. 12.**—The output of our automated line detection procedure for source LAEB21116 showing a typical reliable detection. An emission line is detected with high completeness and reliability $(C, R) = (1.000, 1.000)$ at $\lambda = 4482$ Å. We observe the differences in the top-left, top-right, and middle-left panels from Fig. 11 after the selection procedures. This is a typical example of a reliable detection. See the electronic edition of the PASP for a color version of this figure.
contour significantly overestimates the false detection rate, as expected.

3.2.2. Postprocessing and Reliable Detections

In this section, we describe the second stage selection that results in reliable detections. The bottom panel of Figure 5 shows the R2 contours, completeness contours, and R1 distribution. Since the true reliabilities lie between R1 (conservative false detections) and R2 (generous false detections), we need to set a locus excluding R1 fully and R2 partially. We find that the contour of completeness $= 0.5$ seems to work well, but there is no correct choice. One can choose more generous or strict reliability depending on how the detections will be used. Alternatively, all detections may be used, if weighting by the reliability. In our method, we use this completeness $= 0.5$ line to remove low reliability detections.

The sky residuals are systematic errors resulting from imperfect sky subtraction of strong telluric emission lines and are not

![Detection Histogram](image1)

![Spatial Distribution](image2)

![Pseudoprofile](image3)

![LAEAE15047SPEC](image4)

![Original Profile](image5)

Fig. 13.—The output of our automated line detection procedure for LAEA15047. A single emission line is detected at 4428 Å, but the feature has lower completeness and reliability $(C, R) = (0.777, 0.935)$ than the source in Fig. 12. See the electronic edition of the PASP for a color version of this figure.
governed by gaussian statistics. As a result, our adopted false detection contours are not able to filter out the “detections” resulting from these features. We therefore introduce an additional filtering process in the post processing phase, where we remove any “detections” that lie within ±5 Å of the Hg I 4047 Å and 4358 Å telluric emission lines. If the first detection is excluded by these windows, the second is chosen as the primary detection. If the second detection is also excluded by one of the filters (i.e., by lying inside the false detection contours or too close to a strong telluric emission line), then the third detection is chosen. This case is very rare: for the 1574 spectra, only three cases are found where the third detection is the primary non-rejected detection (§ 3.2.3).

We show three examples of LAEA5569, LAEB21116, and LAEA15047, to illustrate the process of identifying robust detections. In LAEA5569 (shown in Fig. 11), both the first and second detections fall at wavelengths affected by systematic errors due to the subtraction of strong sky lines. While the detections (just barely) lie within our chosen reliability criterion, they are excluded by their wavelength position. In LAEB21116 (Fig. 12), the emission line is well detected, highly reliable, and well separated from any region affected by sky subtraction systematics. In LAEA15047 (Fig. 13), the detected emission line is similar to that in LAEB21116 and is likely to be a reliable detection. However, the feature is less conspicuous in visual inspection, and as the detection measures show lower confidence values of \((C, R) = (0.777, 0.935)\). We refer the reader to the figure captions for detailed descriptions of the characteristic plots produced by our code.

### 3.2.3. Application to Large Samples

We have demonstrated how our method works for data from a single MMT/Hectospec configuration PM1. Now we apply our method to all 7 configurations of the 1574 LAE candidate spectra.

Figure 14 shows the first, second, and third detections from the 1574 spectra in the same format of Figure 9. There are 881 raw detections in the 1574 spectra, 616 of which are first detections and 196 of which are second detections. The raw detection rate of all configurations is 0.56, which is lower than 0.73 of PM1 alone. The drop in the success rate is partly due to weather (the observations in 2012 were taken in poor weather and typically had shorter exposure times) and partly because of changes in the candidate selection between runs (see A. Dey et al. 2015, in preparation for details). As in Figure 9, most of the second and third detections are sky residuals or noise detections that follow the low reliability contours.

Figure 15 shows the final 652 detections from the 881 raw detections. The results shown in the top-left, top-right, and middle-right panels are the same with Figure 12. There is, however, one additional interesting feature revealed by the larger sample.

---

**Fig. 14.** The first, second, and third detections from 1574 spectra in all configurations. The numbers for each detection category are 616 for D100, 196 for D110, and 69 for D111, totaling 881 raw detections out of the 1574 observed candidates. Like Fig. 9, most of the second and third detections are sky residuals or noise detections. See the electronic edition of the PASP for a color version of this figure.
In the top-left panel, we can find the three strips of clustered points. The two horizontal strips at 4047 Å and 4358 Å are produced by the sky residuals as second detections. This is of no interest. The real interesting feature is the ascending strip from 4350 Å to 4550 Å, where the centroids of the first and second detections are similar, $\lambda_{\text{primary}} \approx \lambda_{\text{secondary}}$. This feature results from the detection of double-peaked Ly$\alpha$ emission lines. Although many of the Ly$\alpha$ emitters exhibit single-peaked emission lines (at least, at our spectral resolution of 6 Å), 21 detections exhibit double peaked lines, where both peaks

![Graphs and diagrams showing distributions and detections.](image)
Fig. 16.—The profiles of the final customized detections with C and R values. The left, middle, and right panels show, respectively, the pseudoprofile, the original emission line profile, and the original profile in the full wavelength search range. The detections of $(C, R) \approx (1.0, 1.0)$ are consistent and clear from by-eye inspection. They are also robust to most fluctuations of baseline due to their high S/N. For detections of low reliability and completeness, the signal features are visibly more marginal.
In these cases, some weak detections are disputable as enhanced false detections due to uncertain baselines.

In these cases, some weak detections are disputable as enhanced false detections due to uncertain baselines.
are independently recorded as significant detections by our algorithm.

The bottom-left panel shows the R.A. and decl. positions, the middle-left the R.A. and redshift, and the bottom-right panel the decl. and redshift of the 652 detections. The redshifts are calculated from the observed centroids and the wavelength of Ly$\alpha$ emission. We can find clusters and filaments from the redshift distributions. The heavily populated LAEs near $z \approx 2.67$ form a wall-like structure rather than a compact cluster.

Figure 16 shows 7 randomly chosen LAE detections along with their $C$ and $R$ values. The first column shows pseudoprofiles, the second original profiles, and the third original profiles in the wider spectral range from 4000–5000 Å. Figure 17 shows 32 additional detections with their $C$ and $R$ values in the same format as the third column of Figure 16. Overall the automated detections are consistent with visual inspection. The high significance detections of $C \approx 1.0$ and $R \approx 1.0$ are quite clear and robust signals, while lower $C$ and $R$ detections are indeed more marginal upon visual inspection. For marginal detections, we prefer our automated approach because it allows for greater consistency and the possibility of quantifying the reliability of a detection free from the subjectivity of visual inspection.

4. SUMMARY

We have presented an automated and quantitative method to detect a narrow emission line within a target spectral range. The key point of the method is to generate reliability and completeness contours using Monte Carlo simulations. By comparing the contours and the observed detections, we can assign reliability and completeness values for each detected feature and finalize the detection using customized selection criteria.

We have applied our method to MMT/Hectospec observations of a sample of candidate Ly$\alpha$ emitters. All high significance detections of $(C, R) \approx (1.0, 1.0)$ are clear and robust upon visual inspection. For marginal $(C, R)$ detections, we prefer our automated detection approach, as it is more objective than visual inspection. Though the application presented here is designed to detect single faint emission lines, we can in principle apply our method to more general problems, assuming that the continuum baseline can be properly subtracted. For most cases, this requires continuum fitting using model templates (e.g., Tremonti et al. 2004; Bolton et al. 2012; Lee et al. 2008). Since each kind of object has its own continuum model (synthesized stellar spectra for galaxies, quasar templates for quasars, stellar spectral types for individual stars, dust models for infrared emissions), the continuum subtraction is better left as a separate step performed prior to automated line selection. When combined with a proper method of continuum subtraction, our method can be applied beyond faint emitter detections.

We are grateful to an anonymous referee for comments that have improved this paper. Observations reported here were obtained at the MMT Observatory, a joint facility of the Smithsonian Institution and the University of Arizona. The telescope time was granted both by NOAO, through the Telescope System Instrumentation Program (TSIP; funded by NSF) and through Steward Observatory. This paper uses data products produced by the OIR Telescope Data Center, supported by the Smithsonian Astrophysical Observatory. We are grateful to both the NOAO and Steward Observatory time allocation committees for their continuing support of this program. The research activities of SH and AD are supported by NOAO, which is operated by the Association of Universities for Research in Astronomy under a cooperative agreement with the US National Science Foundation. AD’s research is also supported in part by the Radcliffe Institute for Advanced Study at Harvard University. SH is supported by NASA grants HST GO 10890, GO 11195, and GO 13000. MKMP is supported by a Dark Cosmology Centre Postdoctoral Fellowship; The Dark Cosmology Centre is funded by the Danish National Research Foundation.

REFERENCES

Ahn, C. P., Alexandroff, R., Allende Prieto, C., Anderson, S. F., Anderton, T., Andrews, B. H., et al. 2012, ApJS, 203, 21
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bolton, A. S., Schlegel, D. J., Aubourg, E., Bailey, S., Bhardwaj, V., Brownstein, J. R., et al. 2012, AJ, 144, 144
Dawson, K. S., Schlegel, D. J., Ahn, C. P., Anderson, S. F., Aubourg, E., Bailey, S., et al. 2013, ApJ, 145, 10
Fabricant, D. 2005, PASP, 117, 1411
Jannuzi, B. T., & Dey, A. 1999, in ASP Conf. Ser. 191, Photometric Redshifts and the Detection of High Redshift Galaxies, ed. R. Weymann, et al. (San Francisco: ASP), 111
Kochanek, C. S., et al. 2012, ApJS, 200, 8
Lee, Y. S., Beers, T. C., Sivarani, T., Allende Prieto, C., Koesterke, L., Wilhelm, R., et al. 2008, AJ, 136, 2022
Markwardt, C. B. 2009, in ASP Conf. Ser. 411, Astronomical Data Analysis Software and Systems XVIII, ed. D. Bohlender, D. Durand, & P. Dowler (San Francisco: ASP), 251
Ouchi, M., et al. 2004, ApJ, 611, 660
Prescott, M. K. M., et al. 2008, ApJL, 678, 77
Schlegel, D. J., et al. 2009, in press (arXiv:0904.0468v3)
Tremonti, C. A., et al. 2004, ApJ, 613, 898
Yagi, M., et al. 2002, AJ, 123, 66
Yanny, B., Rockosi, C., Newberg, H. J., Knapp, G. R., Adelman-McCarthy, J. K., Alcorn, B., et al. 2009, AJ, 137, 4377
York, D., et al. 2000, AJ, 120, 1579