Estimating Galaxy Spectral Functions Using Sparse Composite Models

Han Liu*
(Advisors: Larry Wasserman, Christopher Genovese and John Lafferty)
Carnegie Mellon University

January 10, 2007

ABSTRACT

Galaxy spectrometry is an important tool for astronomical research. For a given galaxy spectrum, techniques like wavelet transformation, regression splines, parametric deconvolution, etc., can be applied to estimate the spectral function. However, in many cases, the observed spectrum is the mixture of a smooth continuum and some superimposed spiky lines. The convolution of these two components makes these techniques ineffective. In this paper, we propose a unified regression framework, named **sparse composite model**, for galaxy spectral function estimation. Assuming that all spectral functions can be decomposed into a smooth continuum and the spiky lines, our model represents them using different bases in a sparse manner. This sparse composite model approach has good theoretical guarantees and is very efficient and effective when facing large scale datasets.

Keywords: galaxy spectral function estimation, basis pursuit, matching pursuit, overcomplete dictionary, simultaneous confidence bands, robust estimation

*The preliminary results and the draft version for 36-757: Advanced Data Analysis (I), Fall 2006

I. Introduction

Galaxy spectrometry is an important tool for modern astronomical investigations, enabling astronomers to compare and study galaxy properties in a quantitative manner (Carroll and Ostlie, 1996). A spectrum is a plot of the intensity of the object as a function of wavelength of light. More specifically, a galaxy spectrum is a $x - y$ graph with *wavelength* on the x-axis measured in *Ångstroms* (1Å $= 1 \times 10^{-10}$ meters) and a measure of brightness called *flux* (derived from binned photons counts) on the y-axis. The study of galaxy spectra can reveal information about both continuum processes (e.g. blackbody surface temperature, nonthermal processes) and quantum processes (e.g. absorption and emission lines from electron transitions in heated atoms of various elements).

Each process embodies a corresponding component in the observed spectrum. The smooth continuum component is caused by the combination of a range of blackbody emitters, mainly dense gases or solid objects which radiate heat. The radiation is over a broad range of wavelengths, making the resulting spectrum fairly smooth and continuous. The discrete, non-continuous spiky lines are an observable result of a combination of two sub-processes: the emission and absorption processes. The galaxy emission process is mainly due to gas being heated and then re-radiating energy at specific and distinct wavelength, the fact that each element on the periodic table has its own set of possible energy levels makes the resulted spectrum discrete. If light from a steller core with a continuous spectrum encounters an atom, the wavelengths corresponding to possible energy transitions within the atom will be absorbed. This results in the absorbtion process. If the emission process dominates the absorbtion process, we will observe a very positive spiky line. On the contrary, a negative spiky line is expected. The left subplot of Figure 1 shows a typical galaxy spectrum with the possible spiky line positions superimposed as dashed lines. The right subplot illustrates its variance function at different wavelength, we see that it is highly heteroscedastic.

Galaxy spectra are heavily studied both to understand galaxy properties and to map the luminous mass in the Universe via their redshifts (a surrogate for distance). These properties include the radial velocity of the galaxy (Wakker and van Woerden, 1997), the star-formation rate (Heavens et al., 2004), the kinematics of the galaxy (Saha and Williams, 1994), the average age and metalicity of the stellar populations (Vazdekis and Arimoto, 1999; Moller et al., 1997), etc.. Especially, the continuum and spiky lines tell us completely different stories about the galaxy. The smooth continuum reveals the metallicity of the steller populations, the mass and the radial velocity of the galaxy, etc.. The spiky lines convey the age and temperature information of the steller populations. Due to these fact, a statistically justifiable model which could accurately estimate the smooth continuum and the spiky lines should be significantly helpful for the study of galaxy properties. A reasonable model should satisfy the following properties: (i) The model should be able to estimate the continuum and the spiky lines accurately; (ii) The model should be based on some well-defined statistical principles; (iii) The model should have some physical meanings; (iv) The model should be computationally efficient when facing huge datasets. From a statistician's perspective, each length-n spectrum can be represented as n data points $(x_1, Y_1), ..., (x_n, Y_n)$, where each (x_i, Y_i) represents the observed flux value Y_i at the wavelength x_i, $(i = 1, ..., n)$. Without lose of generality, we assume that x_i is renormalized to lie in the interval $[0, 1]$. Denote $Y = (Y_1, ..., Y_n)$ and $X = (x_1, ..., x_n)$. assuming the observed spectrum value Y_i comes from an unknown spectral function $m(x_i)$ plus some independent measurement error ϵ_i. It's natural to model this problem under a regression framework (Weisberg, 2005):

$$Y_i = m(x_i) + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \quad i = 1, ..., n \tag{1.1}$$

Many existing methods, like wavelet transformation, regression splines, and parametric deconvolution, etc., can solve this regression problem and accurately estimate the function $m(x)$ with the desired properties (ii),(iii), (iv) mentioned above. However, what we are really interested in is not $m(x)$, our task is to accurately estimate the two underlying components that generate $m(x)$, i.e. assuming $m(x) = f(x) + g(x)$, where $f(x)$ represents the continuum component, while $g(x)$ represent the spiky line component, we are interested in finding the underlying f and g based on the observation of Y. Without extra assumptions, simultaneous estimation of both $f(x)$ and $g(x)$ is an ill-posed problem, since if $\widehat{f}(x)$ and $\widehat{g}(x)$ are estimated functions according to a given criterion, then for any constant c, $\widehat{f}(x) + c$ and $\widehat{g}(x) - c$ should also be solutions. This un-uniqueness makes the problem ill-posed. we call it the *decomposition* problem. More related discussion could be found in Li and Speed (2000a,b). The only way to solve this ill-posed problem is by adding more assumptions and constraints. One idea is to use different bases to approximate the continuum component $f(x)$ and the spiky line component $g(x)$, this corresponds to a composite model approach (Sun, 2000). However, the inference of the composite model is based on some heuristic rules, which is hard to justify.

This paper moves towards the solution to these challenges. Our methods are embodies in a unified framework named sparse composite models. The basic idea is that each galaxy spectrum is decomposed into two unknown components: the smooth continuum and the spiky lines. Different components are represented by a linear combination of basis functions from different bases. To tackle the hardness of finding an unique solution in an overcomplete dictionary, we enforce a sparseness constraint, the model can be inferences by basis pursuit or matching pursuit techniques. In statistical terms, this corresponds to "regularize" the estimator. The reason that this regularization technique is helpful to make a better estimation can be traced back to Tikhonov (1963); Chen et al. (1998). Using this approach, the key issues are how to design suitable basis functions for the overcomplete dictionary; how to find a small group of basis functions which could simultaneous represent different galaxy spectra while still achieving sufficient accuracy; how to conduct inference if we do not have any prior information about the spiky line locations; more importantly, the whole work flow should be computationally efficient to make it practically useful. With all these require-

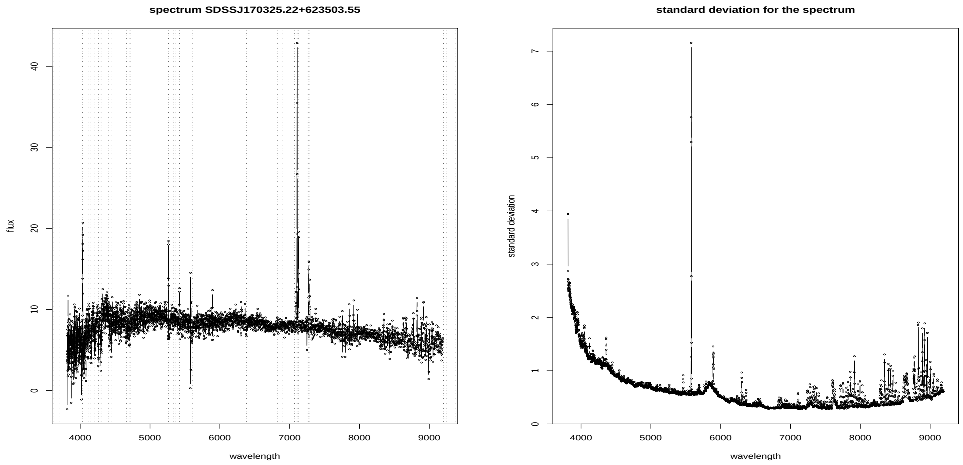


Figure 1: *Left: Plot of a typical galaxy spectrum with possible line positions superimposed. Right: Plot of the standard deviation function at different wavelengths*

ments in mind, we develop a hierarchy of methods to inference sparse composite models: (i) data-adaptive basis pursuit/matching pursuit (DABP/DAMP), (ii) overcomplete dictionary basis pursuit/matching pursuit (ODBP/ODMP), and (iii) simultaneous orthogonal matching pursuit (SOMP). They achieve increasingly stronger conclusions, but at the price of assumptions which are correspondingly more detailed and possibly less reliable. Under the spare composite model framework, simultaneous confidence interval estimation techniques, like, the Scheffé bands and the Tube's bands, can be directly applied. Also, robust versions of different methods using the least absolute deviation Lasso (LAD-Lasso) are also developed, which is expected to better handle the possible heavy-tailed errors or outliers encountered in the galaxy spectra analysis. We test the performance of all these models using both synthetic and real-world datasets and find that ODMP and SOMP outperform the remaining methods, they achieve very good estimate of the spectral functions and the simultaneous confidence bands are very tight. They should be suitable the models which can meet all the above mentioned criteria.

The paper is organized as the following: In section II, we briefly review and compare the previous representation models and summarize the related work. Section III formally presents our sparse composite model framework and the basic ideas of basis/matching pursuit, different theoretical grantees are given in this section. Section IV illustrates five inference methods (DABP, DAMP, ODBP, ODMP, SOMP) and how to design robust basis pursuit method using least absolute deviation Lasso. Section V discusses how to obtain the simultaneous confidence band for the degenerate linear models (e.g. Bonferroni bands, Scheffé bands, Tube's bands, and the bootstrapping bands). Section VI mainly focus on the numerical results of different models from both synthetic and real-world experiments. Section VII gives an overall conclusion and the possible future research directions.

II. Related Work

Many methods have been developed for representing and estimating general spectral functions. According to their modeling assumptions, these methods can be roughly classified into four categories: the *ad hoc* approach (AHA), the *fully parametric* approach (FPA), the *semi-parametric* approach (SPA), and the *linear model* approach (LMA). A brief summarization and comparison of different methods will be given out in this section.

A. Ad hoc Approach

The *ad hoc* approach is currently adopted by the SDSS spectroscopic pipeline[1]. It assumes that all the possible spiky line locations are already known. According to this method, the smooth continuum and the spiky lines are separately fitted in two stages, the procedures of these two stages are based on some intuitive, but fairly heuristic rules : To fit the continuum, a sliding window of length 300 data points is created. Observations closer than 8 data points to any reference line are masked (i.e. not really detectable) and not used in the continuum measurement. The remaining data points are ordered and the values between the 40th and 60th percentile are averaged to give the continuum fitting. After the first stage, spiky line fitting is performed twice in the SDSS spectroscopic pipeline. The first time it is done as part of finding characteristic emission lines in order to measure what is called an emission line *redshift.* Wavelet filters can be used in this step to locate strong emission features in the spectrum. Secondly, after the redshift has been determined, all possible lines are finely searched and accurately measured again. Every line in the reference list is fit as a single narrow Gaussian, say $\mathcal{N}(\mu, \sigma^2)$, on top of the continuum subtracted spectrum. Lines that are deemed close enough are fitted simultaneously as a blend. The basic line fitting is essentially based on the $Levenberg - Marquardt$ nonlinear least square method (Nocedal and Wright, 1999). Parameters are constrained to fall within certain values by multiplying the returned sum of squares by a step function. Generally, the parameter σ is constrained in the interval (0.5Å ,100Å). The final constrained least square problem can be solved using some standard optimization package.

Even though the *ad hoc* method is intuitively sound, it's criticized due to the lack of both physics meaning and statistical guarantees. Especially when our task is trying to provide scientific evidence to some proposed physics theory, this *ad hoc* approach is obvious not suitable due to its lack of justification.

B. Fully Parametric Approach

A fully parametric approach is developed by van Dyk and his Harvard colleagues in a series of papers (Hans and van Dyk, 2003; van Dyk et al., 2001, 2002). Essentially, they use a

[1] see http://www.sdss.org

nonlinear hierarchical Bayesian Poisson regression models to estimate the galaxy spectral functions. They assume that the given spectrum Y has only L levels and is generated from an $L \times 1$ vector of Poisson counts with independent components, i.e.

$$Y_l \sim \text{Poisson}(\lambda_l), \quad l \in \{1, ..., L\} \tag{2.1}$$

where $\Lambda = (\lambda_1, ..., \lambda_L)$ is the vector of expected counts. Λ is modeled in a parametric form

$$\Lambda = \mathbf{PA}\mu + \xi \tag{2.2}$$

where μ is the true function we are interested in, $\mathbf{P}$ is used to model instrument effect, $\mathbf{A}$ is used to model the detector effect, and ξ is an $L \times 1$ vector, used to model the background effect in each level. More detailly, if the observed spectrum is assumed from J bins, because of the instrument effect, p_{lj} is the probability that a photon corresponding to the ideal bin j is recorded by the detector in bin l, thus $\mathbf{P} = \{p_{lj}\}_{l=1,...,L}^{j=1,...,J}$ is the parameter matrix used to model this instrument effect. $\mathbf{A} = \{a_{ij}\}_{i=1,...,J}^{j=1,...,J}$ is a $J \times J$ diagnoal matrix, each element a_{jj} represents the probability that a photon arrives at the detector corresponding to the ideal bin j. The whole model inference is conducted under a hierarchial Bayesian framework.

The most interesting part of this approach is that it incorporates a lot of physics assumptions into parametric statistical models. The model inference is computationally intensive, but is still tractable. However, the success of this approach crucially relies on the validity of the modeling assumptions, which limits the effectiveness of this model. From a statistical point of view, we hope the model can be flexible enough and be more data-driven. This is especially important when we are trying to provide support for some scientific claims.

C. Semiparametric Approach

The semiparametric approach is mainly proposed by Brutti et al. (2005). By their approach, sieved penalized regression spline is used to model the continuum component, while the spiky lines are fitted using a profile likelihood method with each spikes modeled as a narrow Gaussian density function. The whole inference process proceeds in an iterative manner: First, a fully parametric approach is used to get a *pilot* estimate. Then, the estimated continuum is refit using the regression splines. Further, the parameters in the profile likelihood for the spiky lines are re-estimated using the continuum subtracted spectrum. More specifically, a confidence "envelope" around the nonparametric fit is built and then propagate to the parametric component. Therefore, simultaneous confidence bands can be obtained in this manner. This approach is intuitively interesting and computationally efficient. However, it's very hard to analyze the properties of such an estimator due to its iterative algorithm, another problem is that the confidence band generated by this iterative approach tends to be too conservative, see Brutti et al. (2005).

Another interesting semiparametric method is called composite models, which are first proposed by Sun (2000) in the area of of image analysis. The composite model assumes that each spectral function can be decomposed into a smooth component and a spiky component. The Fourier basis are used to model the smooth component and the Dirac δ-function is used to model the spiky lines. The analytical form of the composite model is

$$Y_i = \sum_{j=0}^{m} \left\{ a_j \cos\left(\frac{2\pi j x_i}{n}\right) + b_j \sin\left(\frac{2\pi j x_i}{n}\right) \right\} + \sum_{k=1}^{M} w_k \delta(x_i - l_k) \tag{2.3}$$

where a list of pairs $(l_1, w_1), ..., (l_M, w_M)$ represent the locations and weights of different spiky lines. All these information is known as prior knowledge. Under the assumption that each spectrum contains relatively few spiky lines, the author provides a generic proof that all spectral functions can be represented through a small number of parameters using the composite model. The powerful representation skill of this model in fact comes from the use of an overcomplete dictionary, which is now a very sophisticated topic in the statistics and signal processing communities. The drawback of this composite model approach is that its inference is based on some very heuristic rules, and the locations of the spiky lines must be known as *a priori*. In Sun (2000), the author simply suggests a heuristic value $m = 31$. While the parameters $a_j, b_j, j = 1, ..., m$ are estimated as

$$\widehat{a}_j \;\; = \;\; \frac{2}{n} \sum_{k=1}^{n} Y_k \cos\left(\frac{2\pi j x_k}{n}\right), \;\; \widehat{b}_j = \frac{2}{n} \sum_{k=1}^{n} Y_k \sin\left(\frac{2\pi j x_k}{n}\right), \;\; j = 0, ..., m \tag{2.4}$$

After the continuum is gotten, the spiky lines can be fitted using the continuum subtracted spectrum. In section III, we will illustrate a sparse version of this model, which is more sophisticated and is based on some well studied statistical principles.

D. Linear Model Approach

Comparing with the previous methods, the linear model approach relies on fewer assumptions and tends to be more "nonparametric". It has not been widely used in the galaxy spectra study, but has attracted a lot of attention in the image analysis community due to its flexibilities and solid theoretical foundations (Forsyth, 1990; Drew and Funt, 1992; Marimont and Wandell, 1992; Vrhel et al., 1994).

The key idea of the linear model approach is to augment/replace the wavelength vector $X = (x_1, ..., x_n)$ with p additional variables, denoted as $\phi_i(X)$ $(i = 0, ..., p)$, which are transformations of X, and then fit linear models in this new space of derived input features, that is

$$m(x) = \sum_{i=0}^{p} \beta_i \phi_i(x) \tag{2.5}$$

The coefficients $\beta = (\beta_0, ..., \beta_p)$ is determined by the least square method

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j \phi_j(x_i) \right)^2 \tag{2.6}$$

Two special cases of the linear models are especially widely used in image analysis areas in estimating general spectral functions: the point sampling method and the analytical method.

Point sampling method: If we set $\phi_j(x) = I(x_j - \Delta x/2 \leq x \leq x_j + \Delta x/2)$, this results in the point sampling method. Which represents a spectral function defined on a continuous wavelength region through its functional values at a set of uniformly sampled discrete points. That is, let $m(x)$ be evenly sampled from the range [0,1], the estimated value $\widehat{m}(x)$ is represented as

$$\widehat{m}(x) = Y_i, \quad \text{for} \quad -\frac{\Delta x}{2} < x - x_i < \frac{\Delta x}{2} \tag{2.7}$$

where $i = 1, ..., n$ and $\Delta x = 1/(n-1)$.

This method is quite simple and intuitive, and has been widely used in the area of realistic image synthesis. See Hall (1989); Gondek et al. (1994). However, one severe drawback of the point sampling method is its difficulty in representing spectral functions with spiky lines. This could be well understood with the Shannon's sampling theorem (Marks, 1993), which says that the necessary condition for a function to be perfectly represented by the sampling method is that this function contains no components of frequencies

$$f \geq f_c \equiv \frac{1}{2\Delta x} \tag{2.8}$$

where f_c is called the Nyquist critical frequency and Δx is the sampling interval. Due to the existence of very spiky lines (which corresponds to high frequencies) in the galaxy spectra, the point sampling method is inappropriate in our task.

Analytical method: If we set $\phi_j(x) = x^j$, this is the analytical method (Raso and Fournier, 1991; Geist et al., 1996). The analytical method is essentially a parametric polynomial model. Unfortunately, the computation of polynomial fitting becomes numerically unstable when the polynomial degree is larger than 7 (Forsythe, 1957). This severely limits the application of the polynomial models for spectral function estimation.

The linear model approach has demonstrated advantages in both accuracy and computational efficiency. However, for the purpose of galaxy spectral function estimation, the key issue for the linear model approach is how to design suitable basis functions and how to find a small group of basis functions which could simultaneously approximated all the galaxy spectra in a big dataset. These important problems has not been addressed much. In the next section, we have two models to handle these two problems.

A comparison of different existing approaches are summarized in table 1. The "assumption" item means whether the method relies on very restricted assumption or not. The "Computation" item represents the computational burden of different methods. The "Physics" item shows whether the model incorporates many physical assumptions. The "Stats" item shows whether the model has much statistical meaning or not. The "Spike Pos" item indicate

Table 1: Comparison among different methods

Method	Assumption	Computation	Physics	Stats	Spike Pos	Theory
AHA	very weak	Low	Low	Low	known	×
FPA	strong	medium	High	Low	known	√
SPA	medium	high	Medium	Medium	known	×
LMA	weak	medium	Low	High	known	√

whether the method needs the known spiky line positions as a modeling assumption. The "Theory" item indicates whether the model has some well-justified statistical principles as the theoretical guarantees. Of all these methods, each has different advantages but none meets all the criteria. In comparison with them, the proposed sparse composite model has a better balance among all these criteria.

III. Sparse Composite Models : Main Idea

In this section, we resent the basic concept of the **sparse composite models**. In the next section, we will discuss five methods to fit sparse composite models. These five methods are: (i) overcomplete dictionary basis pursuit/matching pursuit (ODBP/ODMP), (ii) data-adaptive basis pursuit/matching pursuit (DABP/DAMP) and (iii) simultaneous orthogonal matching pursuit (SOMP). The sparse composite models can be classified as linear models, but emphasize more on the design of suitable basis functions for the problem domain of galaxy spectral function estimation. Similar as the linear model method, we first expand the wavelength vector to a bunch of basis functions using some carefully designed criterion. Instead of performing a least square fit like the linear model method, we seek a *sparse* solution to the estimated coefficients. What's more, we are interested in choosing a small group of basis functions which could best represent all the galaxy spectra in a large dataset. Therefore, the spare composite models can be viewed as a hybridization of the basis pursuit or matching pursuit techniques with the composite models. In the following, we will introduce the sparse composite models using an overcomplete dictionary view. Then, basis pursuit, matching pursuit and the uncertainty principles are briefly introduced. Many results in this section can be found in a series of papers by Donoho and his Stanford groups (Donoho, 1992; Donoho and Huo, 2001; Donoho, 2002; Elad and Bruckstein, 2002; Donoho et al., 2006; Donoho and Elad, 2006). Also, a parallel and independent development of these results can be found in Tropp et al. (2003); Tropp (2004, 2005, 2006).

A. Sparse composite models: an overcomplete dictionary view

As is shown in equation (1.1), a galaxy spectrum is a spectral function contaminated with some noise. The composite model approach assume that this spectral function can be decomposed into the form of a smooth continuum and spiky line functions, that is

$$Y_i = f(x_i) + g(x_i) + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{3.1}$$

Under the linear model framework, we want to use basis expansion to estimate the unknown functions $f(x)$ and $g(x)$. However, due to the different characteristics of these two functions (smooth vs. highly spiky). If only one basis is used, for instance, the Fourier base, it's expected to have a poor representation to estimate the highly localized spiky lines. However, if we use the very localized daubechie basis, it's hard to estimate the very smooth continuum. To overcome such difficulties, it's very natural to represent the continuum and the spiky lines using basis functions from different bases. More formally, assuming we use a basis Ψ_f to represent function $f(x)$, every elementary basis function in Ψ_f is globalize and smooth, like Fourier basis. And another basis Ψ_g to represent the spiky line function $g(x)$, the basis functions in Ψ_g are more localized, like Daubechies functions or Dirac δ-function. The combination of these two bases corresponds to an *overcomplete* dictionary $\mathcal{D} = [\Psi_f, \Psi_g]$. Assuming that $\mathcal{D} = (\phi_1, \phi_2, ..., \phi_p)$, it's quite possible that $n < p$. Such an overcomplete representation offers a wider range of generating basis functions, potentially, it's more efficient and flexible at task like estimating the galaxy spectral functions. For this, we want to find a representation for

$$\widehat{m}(x) = \widehat{f}(x) + \widehat{g}(x) = \sum_{i=1}^{p} \widehat{\beta}_i \phi_i(x) = \sum_{\phi_i \in \Psi_f} \widehat{\beta}_i \phi_i(x) + \sum_{\phi_j \in \Psi_g} \widehat{\beta}_j \phi_j(x) \tag{3.2}$$

To find the estimate for $f(x)$ and $g(x)$ separately, it's very natural to set $\widehat{f}(x) = \sum_{\phi_i \in \Psi_f} \widehat{\beta}_i \phi_i(x)$ and $\widehat{g}(x) = \sum_{\phi_j \in \Psi_g} \widehat{\beta}_j \phi_j(x)$. One drawback for this approach is that, due to the fact $n < p$, the representation is undeterminable. From an algebraic view, the solution to equation (3.2) is not unique. Also, the existence of noise in the observed spectrum makes our case worse.

To tackle this ill-posed problem, Donoho showed that if we assume the underlying function $m(x)$ has a sparse representation using the basis functions from the dictionary $\mathcal{D}$ and $\mathcal{D}$ has a property of *mutual incoherence* (below), there exists efficient algorithms to recover the underlying function $m(x)$ uniquely (Donoho et al., 2006). Formally, for the sparse composite models, we are trying to find some coefficient vector β, such that

$$\widehat{\beta} = \arg\min_{\beta} \|\beta\|_0 \quad \text{subject to} \quad \sum_{i=1}^{n} \left(Y_i - \sum_{\phi_j \in \mathcal{D}} \beta_j \phi_j(x_i) \right)^2 \leq \delta \tag{3.3}$$

where $\|\beta\|_0$ is defined to be the number of non-zero elements in β. More details will be introduced in the next subsection.

B. Sparsity, basis pursuit, and matching pursuit

To measure the sparsity of a solution vector β, l_0 norm is adopted, which is simply the number of non-zero elements in β. Assuming the presence of noise, the sparse version of the composite model is defined as

$$(P_{0,\delta}): \quad \widehat{\beta}_{0,\delta} = \arg\min_{\beta} \|\beta\|_0 \text{ subject to } \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j\phi_j(x_i)\right)^2 \leq \delta \tag{3.4}$$

This is a typical combinatorial optimization problem, requiring the enumeration of all possible k-element collections of the basis functions ϕ_j in the dictionary $\mathcal{D}$, for $k = 1, ..., p$, looking for the smallest group which could satisfy the above condition. The time complexity for such an algorithm is in general $O(2^p)$, even if there really exists a sparse k-element representation, the complexity is at least $O(p^k)$. Which is obvious unaffordable by the current digital computers.

Due to the fact that l_0-norm is not a convex function, a natural idea is to relax l_0-norm to l_1-norm, which is convex. By solving the relaxed convex programming problem, we hope the result is very close or exactly the same as the original l_0 problem. This technique is called basis pursuit (BP). It is a convex optimization technique for recovering a sparse signal, based on l_1-norm minimization. The method is due to Chen et al. (1998) and was independently developed by Tibshirani with the name “Lasso” (Tibshirani, 1996). In fact, the basic technique of using an l_1 relaxation to obtain a sparse solution can be traced back to Claerbout and Muir (1973). More detailed historical overview can be found in Tropp (2006). Using this convexification idea, by replacing l_0-norm with l_1-norm, the basis pursuit problem can be defined as

$$(P_{1,\delta}): \quad \widehat{\beta}_{0,\delta} = \arg\min_{\beta} \|\beta\|_1 \text{ subject to } \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j\phi_j(x_i)\right)^2 \leq \delta \tag{3.5}$$

$(P_{1,\delta})$ can be cast as a convex quadratic programming problem and can be solved by many standard optimization procedures, like, the interior-point algorithms (Chen et al., 1998) and the active-set methods, etc.. In statistics, we generally write this problem in the Lasso form, which corresponds the convex optimization in the Lagrangian form

$$(P_{1',\lambda}): \quad \widehat{\beta}^{\lambda} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j\phi_j(x_i)\right)^2 + \lambda\|\beta\|_1 \tag{3.6}$$

for suitable selected $\lambda = \lambda(Y, \delta)$, the solutions of $(P_{1',\lambda})$ and $(P_{1,\delta})$ are exactly the same. Generally, the Lasso problem $(P_{1',\lambda})$ can be solved by an algorithm named LARS, the tuning parameter λ is chosen by the C_p score, for more details, see Efron et al. (2004).

Besides basis pursuit, another relaxation idea is called matching pursuit (MP) (Mallat and Zhang, 1994). Here, we introduce a variant of the MP algorithm, named orthogonal matching pursuit (OMP) due to Tropp and Gilbert (2006). It's a greedy approximation to the problem $(P_{1,\delta})$. The procedure starts from an initial residual $r^{(0)} = Y$ and a current decomposition $\widehat{Y}^{(0)} = 0$; then for $k = 1, ...$, it augments the decomposition from $\widehat{Y}^{(k-1)}$ to $\widehat{Y}^{(k)}$ and updates the residual $\widehat{r}^{(k-1)}$ to $\widehat{r}^{(k)}$ in a stepwise manner, always maintaining $Y = \widehat{Y}^{(k)} + \widehat{r}^{(k)}$. More detaily, assuming all the basis functions from the dictionary $\mathcal{D}$ are normalized, i.e. $\|\phi_j\|_2 = 1$. At the k-th stage, the matching pursuit algorithm selects an atom to be added to the decomposition based on correlation with the current residual

$$i_k = \arg\max_{1\leq i\leq p} |\langle r^{(k-1)}, \phi_i\rangle| \tag{3.7}$$

The estimated value at the step k is represented as

$$\widehat{Y}_j^{(k)} = \sum_{l=1}^{k} \widehat{\beta}_{i_l}^{(k)} \phi_{i_l}(x_j), \quad j = 1, ..., n \tag{3.8}$$

where the coefficients $\widehat{\beta}_{i_l}^{(k)}$ are fitted by the ordinary least squared regression to minimize the residual $r^{(k)}$. The estimated residual will be the input for the next step. The algorithm proceeds until l_2-norm of the residual is less than δ. In the next subsection, we will see that even basis pursuit and matching pursuit are greedy approximation algorithms, underly fairly reasonable assumptions, they can exactly solve the problem $(P_{0,\delta})$.

C. Stable recovery of the basis pursuit and matching pursuit in the presence of noise

As shown in the previous subsection, both BP and MP are much more practical approximate algorithms used to solve problem $(P_{0,\delta})$. However, under some conditions, theory has been developed that BP and MP can correctly solve the problem $(P_{0,\delta})$ up to a constant proportion to the noise level. The concept of *mutual coherence* of the dictionary $\mathcal{D}$ plays a key role in all the theorems. It's defined as

Definition 3.1. *assuming the columns of dictionary $\mathcal{D}$ are normalized to unit l_2-norm, define the Gram matrix $\mathbf{G} = \mathcal{D}^T\mathcal{D}$. With $G(k, j)$ denoting entries of this matrix, the mutual coherence is defined as*

$$M = M(\mathcal{D}) = \max_{1\leq k,j\leq p, k\neq j} |G(k, j)| \tag{3.9}$$

A dictionary is incoherent if M is small. Assume that the true function $m(x) = \sum_{i=1}^{p} \beta_i^{(0)} \phi_i(x)$ and we observed the noisy observations $Y_i = m(x_i) + \epsilon_i$, where $\sum_{i=1}^{n} \epsilon_i^2 \leq \delta$. If $(P_{0,\delta})$ is applied, the following theoretical guarantees can be gotten (Donoho et al., 2006)

Theorem 3.2. *Let the dictionary $\mathcal{D}$ have mutual coherence $M = M(\mathcal{D})$. Suppose the noiseless signal $m(x) = \sum_{i=1}^{p} \beta_i^{(0)} \phi_i(x)$, where $\beta^{(0)}$ satisifes*

$$\|\beta^{(0)}\|_0 = N < \frac{1/M + 1}{2} \tag{3.10}$$

Then, β^0 is the unique sparest such representation of $m(x)$; and the reconstruction $\widehat{\beta}^{(0,\delta)}$ from applying $(P_{0,\delta})$ to the noisy data Y approximates $\beta^{(0)}$:

$$\left\|\widehat{\beta}^{(0,\delta)} - \beta^{(0)}\right\|_2^2 \leq \frac{4\delta^2}{1 - M(2N-1)} \tag{3.11}$$

The above theorem shows that provided that the underlying object has a sparse representation in the dictionary $\mathcal{D}$ and if $\mathcal{D}$ has a low mutual coherence, recovery by explicitly imposing the sparsity yields an approximation to the ideal sparse decomposition of the noiseless signal in which the error is at worst proportional to the input noise level. A parallel development of this result for the basis pursuit algorithm is shown in the following theorem.

Theorem 3.3. *Let the overcomplete dictionary $\mathcal{D}$ have mutual coherence $M = M(\mathcal{D})$. Suppose the noiseless signal $m(x) = \sum_{i=1}^{p} \beta_i^{(0)} \phi_i(x)$, where $\beta^{(0)}$ satisifes*

$$\|\beta^{(0)}\|_0 = N < \frac{1/M + 1}{4} \tag{3.12}$$

Then, β^0 is the unique sparest such representation of $m(x)$; moreover, the solution $\widehat{\beta}^{(0,\delta)}$ from applying $(P_{1,\delta})$ to the noisy data Y approximates $\beta^{(0)}$:

$$\left\|\widehat{\beta}^{(1,\delta)} - \beta^{(0)}\right\|_2^2 \leq \frac{4\delta^2}{1 - M(4N-1)} \tag{3.13}$$

The proof of theorem 3.3 relies on a series of relaxations, each one expanding the feasible set and increasing the maximal value. Therefore, the resulted bound might not be tight, however, the above bound is already quite reasonable and is enough to demonstrate the power of BP. The $(P_{0,\delta})$ and $(P_{1,\delta})$ are two global optimization algorithms, while the OMP is a greedy stepwise approximation algorithm. Paralleling this distinction, only a local stability result can be developed for OMP.

Assuming that the order of basis functions $\phi_1, \phi_2 ..., \phi_p$ in the overcomplete dictionary $\mathcal{D}$ has been chosen so that $m(x) = \sum_{i=1}^{p} \beta_i^{(0)} \phi_i(x)$, the entries in $\beta^{(0)}$ are arranged in a decreasing order and the first N entries are non-zero.

Theorem 3.4. *Suppose the ideal noiseless signal $m(x)$ has a representation $m(x) = \sum_{i=1}^{p} \beta_i^{(0)} \phi_i(x)$ satisfying*

$$N = \|\beta^{(0)}\|_0 \leq \frac{1+M}{2M} - \frac{1}{M} \cdot \frac{\delta}{\beta_N^{(0)}} \tag{3.14}$$

Then $\beta^{(0)}$ is the unique sparest representation of $m(x)$. Denote by $\widehat{\beta}^{(O,\delta)}$ the result of greedy stepwise least-squares fitting which stops as soon as the representation error $\leq \delta$. Then, $\widehat{\beta}^{(O,\delta)}$ has the correct sparsity pattern:

$$\mathrm{supp}(\widehat{\beta}^{(O,\delta)}) = \mathrm{supp}(\beta^{(0)}) \tag{3.15}$$

Also, $\widehat{\beta}^{(O,\delta)}$ approximates the ideal noiseless representation:

$$\left\|\widehat{\beta}^{(O,\delta)} - \beta^{(0)}\right\|_2^2 \leq \frac{\delta^2}{1 - M(N-1)} \tag{3.16}$$

The result shown in theorem 3.4 is only a local stability property, since it's only valid for sufficient small $\delta < \delta^*(\beta^{(0)})$. This theorem also proved the "support" property of the OMP algorithm, i.e. the support of the estimated coefficients by the OMP algorithm is the same as the support estimated by the support of the true spare coefficient vector $\beta^{(0)}$. A similar support property is also derived for the BP approach, as is shown in the next theorem.

Theorem 3.5. *Suppose that $Y_i = m(x_i) + \epsilon_i$ where $m(x) = \sum_{i=1}^{p} \beta_i^{(0)} \phi_i(x)$, $\|\beta^{(0)}\|_0 \leq N$ and $\sum_{i=1}^{n} \epsilon_i^2 \leq \delta$. Let $M = M(\mathcal{D})$ and suppose $\alpha \equiv MN < 1/2$. Set*

$$\gamma = \sqrt{\frac{1-\alpha}{1-2\alpha}} \tag{3.17}$$

solve $(P_{1,\delta'})$ with exaggerated noise level $\delta' = C \cdot \delta$, where $C = C(M,N) = \gamma\sqrt{N}$. Then $\mathrm{supp}(\widehat{\beta}^{(1,\delta')}) \subset \mathrm{supp}(\beta^{(0)})$.

The support property provided by theorem 3.5 is fairly conservative, it says that if we solve the l_1 minimization BP problem with an exaggeration of the noise level, the solution $\widehat{\beta}^{(1,\delta)}$ has its support contained in the support of the true coefficient vector $\beta^{(0)}$.

All the results in this section will be used as guidelines and theoretical guarantees in the next section to design different inference algorithms for the sparse composite models.

IV. Methodology

With the basic idea of sparse composite models, we develop five inference methods to estimate the galaxy spectral functions. They are,(i) data-adaptive basis pursuit/matching pursuit (DABP/DAMP), (ii) overcomplete dictionary basis pursuit/matching pursuit (ODBP / ODMP), and (iii) simultaneous orthogonal matching pursuit (SOMP), using more and more restricted assumptions. The difference between these five methods mainly lies in how to design a suitable overcomplete dictionary, and how to conduct the basis pursuit or matching pursuit to select the most suitable basis functions. To use the ODBP method, the reference spiky line locations is known as prior knowledge, but the ODMP, DABP and DAMP methods do not have this constraint. All of them are computationally efficient and has good scalability when facing large datasets. Under this framework, simultaneous confidence bands for the linear models can be directly adopted. In the following part of this section, we will introduce these three frameworks in more detail. Finally, in this section, we also introduce the idea of robust basis pursuit, which is expected to achieve a better performance when facing outliers or heavy tailed distributions.

A. Overcomplete dictionary basis pursuit/matching pursuit

The idea of overcomplete dictionary basis pursuit/matching pursuit (ODBP/ODMP) is to use different fixed, data-independent bases to represent the smooth continuum and the spiky lines separately. In our case, since the continuum function $f(x)$ is very smooth, Fourier basis is used to construct Ψ_f. Given a wavelength vector X with length n, the discrete Fourier basis (Fast Fourier Transform) employs a set of orthonormal periodic functions, they are in fact a combination of sine and cosine functions. Assuming there are altogether $2m+1$ basis functions in the basis $\Psi_f = [\phi_0, \phi_1, ..., \phi_{2m}]$, they are

$$\begin{aligned} \phi_0(x_i) &= \frac{1}{\sqrt{n}}, \quad i = 1, ..., n \\ \phi_k(x_i) = \frac{\cos(kx_i)}{C_k}, \quad \phi_{k+1}(x_i) &= \frac{\sin((k+1)x_i)}{C_{k+1}}, \text{ where } \ k = 1, 3, ..., 2m-1 \end{aligned} \tag{4.1}$$

where C_k, $k = 1, ..., 2m$ are normalization constants to make sure l_2-norm of $\phi_k(x)$ equals 1. To the purpose of representing the spiky spikes, a very natural choice is to use the Dirac delta function, often referred to as the unit impulse function to represent the spiky lines. There are two possible approaches to construct the base Ψ_g, one approach is to construct a basis with exact n basis functions, i.e. $\Psi_g = [\psi_1, ..., \psi_n]$, each ψ_i is defined as

$$\psi_i(x) = \delta(x - x_i), \quad \text{where} \ \ i = 1, .., n \tag{4.2}$$

Using this approach, the overcomplete dictionary $\mathcal{D} = [\Psi_f, \Psi_g]$ will contain $n + 2m + 1$ basis functions. Generally, for the galaxy spectral function estimation problem, $n \approx 4000$, we can not afford basis pursuit approach due to this very large scale quadratic programming problem. However, matching pursuit can still be applied in this case and is still efficient. This approach is called overcomplete dictionary matching pursuit (ODMP).

If we have the reference spiky line positions as a prior knowledge, the size of the base Ψ_g can be significantly trimmed. Due to the fact that some peaks are wide and some peaks are very sharp, the area covered by a peak has very interesting physics meaning. Therefore, given a reference spiky line position x_i, when constructing the basis functions for the Ψ_g, all the points within a neighborhood of size 16 are also considered as possible spike positions. More formally, assuming a reference spiky line location is x_r. Then 16 basis functions are constructed as

$$\psi_k(x) = \delta(x - x_k), \quad \text{subject to} \ \ x_k \in \mathcal{V}_{16}(x_r) \tag{4.3}$$

where $\mathcal{V}_{16}(x)$ represents the 16-neighborhood of a point x. If a point is in the neighborhood of two different reference points, we only count it once. Assuming there are altogether l

reference spiky line positions, using this approach, we see that the size of the base Ψ_g can at most be $16l$. Since $l \ll n$, we can afford the computational burden of the basis pursuit now. This approach is called overcomplete dictionary basis pursuit (ODBP).

Since both Fourier base and the Dirac's delta function base are orthonormal bases. The correlation between a *cos* function and the δ function is very small, which implies a very low mutural coherence of the overcomplete dictionary $\mathcal{D}$. According to the previous theorems 3.3 and 3.4. Both basis pursuit and matching pursuit should work in this case. The size of the base Ψ_f is a tuning parameter, generally, we choose m as large as possible subject to a small mutual coherence value. In the next section, we will show that the choice of m is not quite sensitive to the algorithm.

One thing to note is that, for the DABP method, instead of using the standard basis pursuit or Lasso, we use a trick named *relaxed Lasso*, which is developed by Meinshausen (2005). Relaxed Lasso is a generalization of the Lasso shrinkage technique for linear regression. The results include all standard Lasso solutions but allow often for sparser models while having similar or even slightly better predictive performance if many predictor variables are present. Using the Lasso solution in equation 3.6, the relaxed Lasso solution is defined as the following:

Definition 4.1. *Assuming the set of predictor variables selected by the Lasso estimator* $\widehat{\beta}^{\lambda}$ *as* $\mathcal{M}_{\lambda}$

$$\mathcal{M}_{\lambda} = \{1 \leq k \leq p | \widehat{\beta}^{\lambda}_k \neq 0\} \tag{4.4}$$

For some parameter $\gamma \in [0, 1]$, *the relaxed Lasso estimator is defined as*

$$\widehat{\beta}^{\lambda,\gamma} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{k \in \mathcal{M}_{\lambda}} \beta_k \phi_k(x_i) \right)^2 + \gamma\lambda \|\beta\|_1 \tag{4.5}$$

where $\gamma\lambda$ *together as new tuning parameter is chosen by the* C_p *score.*

Under some regularity assumptions, this relaxed Lasso estimator is expected to have a better convergence rate than the ordinary Lasso estimator, see Meinshausen (2005). Since relaxed Lasso can get a sparser solution than the ordinary Lasso and still have the same theoretical guarantees, in the sequel of this paper, whenever we mention the basis pursuit or Lasso algorithm, we are in fact refer to this relaxed version.

B. Data-adaptive basis pursuit/maching pursuit

Data-adaptive basis pursuit or matching pursuit algorithms (DABP/DAMP) are motivated from a nonparametric inference technique named Rodeo (Lafferty and Wasserman, 2005). Unlike the commonly used data-independent basis (e.g. Fourier basis, Haar basis or radial basis), data-adaptive basis pursuit/matching pursuit does not need an explicit parametric form of each basis function. This purely data-driven basis function has essentially the same distribution features as the data, and is expected to be more efficient in coding the unknown function $m(x)$.

The basic idea of the data-adaptive basis pursuit is as the following: fix a wavelength x and let $\widehat{m}_h(x)$ denote a nonparametric estimator of $m(x)$ with a smoothing parameter h (e.g. If we we use the univariate local linear smoother, h is the bandwidth parameter). Let $M(h) = \mathbb{E}(\widehat{m}_h(x))$ denote the mean of $\widehat{m}_h(x)$. Assume that $x = x_i$ is a given wavelength value and that $\widehat{m}_0(x) = Y_i$ is an observed function value. In this case, $m(x) = M(0) = \mathbb{E}(Y_i)$. If $\mathcal{H} = \{h(t) : 0 \leq t \leq 1\}$ is a smooth path through the set of smoothing parameters with $h(0) = 0$ and $h(1) = 1$, then

$$\begin{aligned} m(x) &= M(0) = M(1) + M(0) - M(1) && (4.6)\\ &= M(1) - \int_0^1 \frac{dM(h(s))}{ds} ds = M(1) - \int_0^1 \frac{dM(h)}{dh} \cdot \frac{dh(s)}{ds} ds && (4.7)\end{aligned}$$

The last equation follows from the chain rule, a naive estimator

$$\widehat{m}(x) = \widehat{m}_1(x) - \int_0^1 \frac{d\widehat{m}_h(x)}{dh} \cdot \frac{dh(s)}{ds} ds \tag{4.8}$$

is identically equal to $\widehat{m}_0(x) = Y_i$, which has poor risk since the variance of $\frac{d\widehat{m}_h(x)}{dh}$ is large for small h. We need to replace it by an estimate $\widehat{D}(x, h)$ which can be a shrinkage or thresholding estimator of the derivative to reduce the variance. Our estimate of $m(x)$ is then

$$\widehat{m}(x) = \widehat{m}_1(x) - \int_0^1 \widehat{D}(x, h(s)) \cdot \frac{dh(s)}{ds} ds \tag{4.9}$$

Assuming $0 = s_0 < s_1 < \cdots < s_J = 1$, the discrete approximation to this is

$$\widetilde{m}(x) = \widehat{m}_1(x) - \sum_{j=0}^{J-1} \widehat{D}(x, h(s_j)) \cdot (h(s_{j+1}) - h(s_j)) \tag{4.10}$$

To implement this idea, for $j = 0, ..., J-1$, let

$$\phi_j(x_i) = \left(\frac{d\widehat{m}_{h(s_j)}(x_i)}{dh(s_j)} \right) \cdot (h(s_{j+1}) - h(s_j)) \tag{4.11}$$

To reduce the variance of this estimates, we apply soft-thresholding on the obtained basis basis function $\phi_j(x)$. Soft-thresholding shrinks the obtained basis functions towards 0, hoping to prove the estimation performance, more details can be found in Donoho and Johnstone (1994), the rule is

$$\phi'_j(x) = \text{sign}(\phi_j(x))(|\phi_j(x) - K_j|)_+ \tag{4.12}$$

where $K_j = \widehat{\sigma}_j \sqrt{\log n/n}$, and $\widehat{\sigma}$ represents the median absolute deviation (MAD) of the difference vector of $(\phi_j(x_1), ..., \phi_j(x_n))$. To calculate it, we first define a length $n-1$ vector, $\psi = (\psi_1, ..., \psi_{n-1})$, where

$$\psi_i^{(j)} = \phi_j(x_{i+1}) - \phi_j(x_i), \quad i = 1, ..., n-1 \tag{4.13}$$

then $\widehat{\sigma}_j$ is defined as

$$\widehat{\sigma}_j = \frac{\text{median}_i |\psi_i^{(j)} - \text{median}_k \psi_k^{(j)}|}{0.6745} \tag{4.14}$$

This corresponds to a universal soft-thresholding rule, it's not data-adaptive. There is also a more sophisticated data-adaptive soft-thresholding method based on the empirical Bayes approach, see Johnstone and Silverman (2004).

For the implementation purpose, we first fit a sequence of local polynomial regression functions indexed by a sequence of decreasing bandwidths, each basis function is calculated as the difference between two adjacent estimated regression lines in a greedy manner. The resulting basis are further selected by basis pursuit (DABP) or matching pursuit (DAMP), and the final estimation is obtained by fitting a linear model with ordinary least squares. More specifically, to construct the data-adaptive basis function $\phi_j(x_i)$, we first build a sequence of decreasing bandwidths

$$\mathcal{H} = \{h_j : h_j = h_0 \beta^j \text{ for } , 0 < \beta < 1, j = 0, ..., J-1\} \tag{4.15}$$

where $h_0 = c$ is a large enough constant. The data-adaptive basis are defined as the following:

$$\begin{aligned} \phi_0(x_i) &= \widehat{m}_{h_0}(x_i) && (4.16) \\ \phi_j(x_i) &= \widehat{m}_{h_j}(x_i) - \widehat{m}_{h_{j-1}}(x_i), \text{ for } j \geq 1 && (4.17) \end{aligned}$$

where $\widehat{m}_h(x)$ is the local polynomial estimator at point x with bandwidth h. Then, for each obtained ϕ_j, $j = 0, ..., J-1$, apply the thresholding strategy as in formula 4.12 to get ϕ'_j.

With the obtained dictionary $\mathcal{D} = \{\phi'_j(\cdot)\}_{j=0}^{J-1}$, for DABP, we solve a relaxed Lasso problem

$$\widehat{\beta}^{\lambda} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j \phi'_j(x_i) \right)^2 + \lambda \|\beta\|_1 \tag{4.18}$$

Here, $\lambda \in [0, \infty)$ is a tuning parameter chosen by the minimal C_p score. By defining $\mathcal{M}_\lambda = \{1 \leq k \leq p | \widehat{\beta}_k^{\lambda} \neq 0\}$, the relaxed Lasso estimator is

$$\widehat{\beta}^{\lambda,\gamma} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{k \in \mathcal{M}_\lambda} \beta_k \phi'_k(x_i) \right)^2 + \gamma\lambda \|\beta\|_1 \tag{4.19}$$

where the tuning parameter $\gamma\lambda$ is chosen by the minimal C_p score. For the DAMP, we just apply the OMP algorithm as is shown in the previous section.

The dictionary $\mathcal{D}$ generated by this approach is obvious not orthogonal. But all the theoretical guarantees in the previous still hold in this case. A very high level rationale about why this data adaptive approach works is that, at the very beginning of the algorithm, very large bandwidth parameter is used, under the assumption that there are only relatively few spiky lines in the spectrum, the fitting of the local linear smoother is not affected by the spiky lines (i.e. the fitting is mainly based on the smooth continuum). The basis functions generated by the difference between the local linear estimator fits implicitly construct the base Ψ_f. With the proceed of the algorithm, the bandwidth becomes smaller and smaller, the local linear fit will be more and more dominated by the spiky lines. In this case, the bassi functions generated in this stage implicitly construct the base Ψ_g. When constructing the overcomplete dictionary $\mathcal{D} = [\Psi_f, \Psi_g]$, there is not a borderline clearly distinguish Ψ_f and Ψ_g. For this, an empirical hypothesis testing technique will be designed and be tested in the experiments section. When constructing the overcomplete dictionary, an important tuning parameter is β in formula 4.15. A very simple way is to try a sequence of β values and choose the one that minimize the mutual coherence parameter of the overcomplete dictionary $\mathcal{D}$. More detailed will be discussed in the experiment section.

C. Simultaneous orthogonal matching pursuit

Simultaneous orthogonal matching pursuit (SOMP) is a generalization of the orthogonal matching pursuit algorithm (OMP). It's developed by Tropp et al. (2006a,b). SOMP can find a good approximation of several galaxy spectra at once using different linear combinations of the same elementary basis functions. At the same time, it tries to balance the error in approximation against the total number of elementary basis functions that participate. These representative basis functions form a parse representative group for all the input galaxy spectra, they are chosen from a large, overcomplete dictionary $\mathcal{D}$. Before the description of the SOMP algorithm, we first introduce some notations. Assuming that we have altogether K galaxy spectra measured on the same wavelength grid X, the i-th spectrum is represented as $Y^{(i)} \in \mathbb{R}^n$, the $n \times K$ spectra matrix $\mathbf{S}$ is defined as

$$\mathbf{S} = [Y^{(1)}, ..., Y^{(k)}] \tag{4.20}$$

Assuming we have an overcomplete dictionary $\mathcal{D}$, which is a n by p matrix constructed using the data-independent basis functions or data-adaptive basis functions. The *coefficient matrix*, denoted as $\mathbf{B}$, is an element of the linear space $\mathbb{R}^{p \times K}$. The (j, k) entry of the coefficient matrix is written as $\beta_j^{(k)}$ or in functional notation as $\mathbf{B}_{jk}$. From which ,we see that, given a coefficient matrix $\mathbf{B}$, the matrix product $\mathbf{S} = \mathcal{D}\mathbf{B}$ yields a signal matrix. Also, it's obvious that

$$Y^{(k)} = \mathcal{D}\beta^{(k)} = \sum_{j=1}^{p} \beta_j^{(k)} \phi_j \tag{4.21}$$

Using these notations, a formal description of the algorithm is shown in figure 2

Simultaneous orthogonal matching pursuit: algorithm

1. INPUT: An $n \times K$ spectra matrix $\mathbf{S}$ and a stopping criterion
2. OUTPUT:
   (a) A set Λ_T containing T indices, where T is the number of iterations completed
   (b) An $n \times K$ approximation matrix $\mathbf{A}_T$ and an $n \times K$ residual matrix $\mathbf{R}_T$
3. INITIALIZE: $\mathbf{R}_0 = \mathbf{S}, \Lambda_0 = \emptyset$ and the iteration counter $t = 1$
4. LOOP over t, until the stopping criterion is met
   (a) Find an index λ_t that solves the easy optimization problem

   $$\lambda_t = \arg \max_{j \in \{1,\dots,p\}} \sum_{k=1}^{K} |\langle \mathbf{R}_{t-1} e_k, \phi_j \rangle| \tag{4.22}$$

   We use e_k to denote the k-th canonical basis vector in $\mathbb{R}^K$
   (b) Set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$
   (c) Determine the orthogonal projector $\mathbf{P}_t$ onto the span of the basis functions indexed in Λ_t
   (d) Calculate the new approximation and residual:

   $$\mathbf{A}_t = \mathbf{P}_t \mathbf{S} \quad \mathbf{R}_t = \mathbf{S} - \mathbf{A}_t \tag{4.23}$$

Figure 2: The SOMP algorithm.

There are different possibilities for the stopping criterion of the SOMP algorithm. We can simply stop the algorithm after a fixed number T of iterations, or wait until the Frobenius norm of the residual declines to a threshold δ, i.e. $\|\mathbf{R}_t\|_F \leq \delta$. Even though SOMP is a greedy algorithm, good theoretical properties can be developed, see Tropp et al. (2006a).

The development of these theories depends on a newly defined function, named *cumulative coherence function* $\mu(t)$, which is a generalization of the previously defined mutual coherence $M(\mathcal{D})$. It's defined as the following:

Definition 4.2. *For each natural number t, the mutual coherence function $\mu(t)$ is defined* *as*

$$\mu(t) \equiv \max_{\Lambda \subset \{1,...,p\}} \max_{j \notin \Lambda} \sum_{k \in \Lambda} |\langle \phi_j, \phi_k \rangle| \tag{4.24}$$

where the index set $\Lambda \subset \{1, ..., p\}$.

Some definitions of the vector and matrix norms will be used in the following theorems. First, the *row support* of a coefficient matrix is defined as the set of indices for its nonzero rows. More precisely,

$$\text{rowsupp}(\mathbf{C}) \equiv \{j \in \{1, ..., p\} : \beta_j^{(k)} \neq 0 \text{ for some } k\} \tag{4.25}$$

That is, the support of a coefficient vector is the set of indices at which it is nonzero. Further, the row-l_0 quasi-norm of a coefficient matrix is defined as the number of nonzero rows

$$\|\mathbf{C}\|_{\text{row}-0} \equiv |\text{rowsupp}(\mathbf{C})| \tag{4.26}$$

Another useful, but more complicated definition is the concept of *operator norm.* If $\mathbf{A}$ is a matrix with appropriate dimensions, we may view it as linear operator acting on X via left matrix multiplication to produce elements of Y. We call $\mathbf{A}$ maps X to Y. Formally, the adjoint $\mathbf{A}^*$ is treated as a map between the dual space Y^* and X^*. In the current setting, $\mathbf{A}^*$ is simply the conjugate transpose of $\mathbf{A}$, and it also acts by left matrix multiplication.

Definition 4.3. *If* $\mathbf{A}$ *maps* X *to* Y*, its operator norm is defined as*

$$\|A\|_{X,Y} \equiv \sup_{x \neq 0} \frac{\|\mathbf{A}x\|_Y}{\|x\|_X} \tag{4.27}$$

To derive the theoretical guarantees for the SOMP algorithm, an ideal simultaneous sparse matching pursuit problem is defined

$$(\text{SPARSE}): \quad \mathbf{C}_{\text{opt}} = \arg \min_{\mathbf{C} \in \mathbb{R}^{p \times K}} \|\mathbf{S} - \mathcal{D}\mathbf{C}\|_F \quad \text{subject to} \quad \|\mathbf{C}\|_{\text{row}-0} \leq T \tag{4.28}$$

Then, according to Tropp et al. (2006a)

Theorem 4.4. (SOMP with a sparsity bound) . *Assume that* $\mu(T) \leq \frac{1}{2}$*. Given an input matrix* $\mathbf{S}$*, suppose that* $\mathbf{C}_{\text{opt}}$ *solves (*SPARSE*) and that* $\mathbf{A}_{\text{opt}} = \mathcal{D}\mathbf{C}_{\text{opt}}$*. After* T *iterations, SOMP will produce an approximation* $\mathbf{A}_T$ *that satisfies the error bound*

$$\|\mathbf{S} - \mathbf{A}_T\|_F \leq \left[1 + KT \frac{1 - \mu(T)}{(1 - 2\mu(T))^2}\right]^{1/2} \cdot \|\mathbf{S} - \mathbf{A}_{\text{opt}}\|_F \tag{4.29}$$

*In words, SOMP is an approximation algorithm for (*SPARSE*)*

This theorem shows that the error in the computed approximation is never more than a constant factor greater than the optimal approximation error. One drawback of this theorem is that it seems too pessimistic. The factor of T appears in the constant seems an artifact of the proof method. The real performance of SOMP is expected to be much better than this bound. Also, a theorem about the support property of the SOMP algorithm is derived

Theorem 4.5. (SOMP with an error bound) . *Let* Λ_{opt} *be an index set containing* T *basis functions or fewer, where* $\mu(T) \leq \frac{1}{2}$. *Given an input matrix* $\mathbf{S}$*, suppose that* $\mathbf{C}_{\text{opt}}$ *solves* (SPARSE) *and that* $\mathbf{A}_{\text{opt}} = \mathcal{D}\mathbf{C}_{\text{opt}}$. *While it satisfies an error bound*

$$\|\mathbf{S} - \mathbf{A}_{\text{opt}}\|_F \leq \delta \tag{4.30}$$

Let SOMP stop at the end of iteration t*, if the norm of the residual satisfies*

$$\|\mathbf{S} - \mathbf{A}_t\|_F \leq \left[1 + KT\frac{1-\mu(T)}{(1-2\mu(T))^2}\right]^{1/2} \cdot \delta \tag{4.31}$$

It follows that each atom chosen is optimal, i.e. $\Lambda_t \subset \Lambda_{\text{opt}}$.

This theorem states that SOMP can calculate an approximation that achieves an error within a constant factor of δ. Meanwhile, it guarantees that every basis function selected in the computed approximation is drawn from the ideal set of basis functions.

When using SOMP, it's often possible to develop estimates on the correlation between the dicionary $\mathcal{D}$ and the residual left over in approximation. The bounds on this correlation are also available as a theoretical result of SOMP. For this, first define a quantity

$$L(t) = L(t;T) \equiv \frac{\mu(T-t)}{1-\mu(t)} \quad \text{for} \quad t = 0, ..., T \tag{4.32}$$

Then, the following theorem holds

Theorem 4.6. (SOMP with a correlation bound). *Suppose that* Λ_{opt} *lists at most* T *basis functions, where* $\mu(t) < \frac{1}{2}$ *and* $L(T) < \frac{1}{2}$. *Let* $\mathbf{S}$ *be a spectra matrix,* $\mathbf{A}_{\text{opt}}$ *its best approximation over* Λ_{opt}*, and* $\mathbf{C}_{\text{opt}}$ *be the coefficient matrix that synthesizes* $\mathbf{A}_{\text{opt}}$. *Finally, assume we have a bound*

$$\|\mathcal{D}^*(\mathbf{S} - \mathbf{A}_{\text{opt}})\|_{\infty,\infty} \leq \tau \tag{4.33}$$

After iteration t *of SOMP, halt the algortihm if*

$$\|\mathcal{D}^*\mathbf{R}_t\|_{\infty,\infty} \leq \frac{1-L(t)}{1-2L(t)} \cdot \tau \tag{4.34}$$

If the algorithm terminates ate the end of iteration t*, we may conscious that*

1. *the algorithm has chosen t indices from Λ_{opt}, and*

2. *it has identified every index λ from Λ_{opt} for which*

$$\sum_{k=1}^{K} |\mathbf{C}_{\text{opt}}(\lambda, k)| > \frac{\tau}{1 - 2L(t)} \tag{4.35}$$

3. *The absolute error in the computed approximation satisfies*

$$\|\mathbf{S} - \mathbf{A}_t\|_F^2 \leq \|\mathbf{S} - \mathbf{A}_{\text{opt}}\|_F^2 + \tau^2 \left[\frac{1 - L(t)}{1 - 2L(t)}\right]^2 \frac{T - t}{1 - \mu(T - t)} \tag{4.36}$$

4. *In particular, if*

$$\min_{\lambda \in \Lambda_{\text{opt}}} \sum_{k=1}^{K} |\mathbf{C}_{\text{opt}}(\lambda, k)| > \frac{\tau}{1 - 2L(t)} \tag{4.37}$$

   then $t = T$, $\Lambda_t = \Lambda_{\text{opt}}$, and $\mathbf{A}_t = \mathbf{A}_{\text{opt}}$

In summary, theorem 4.6 says that SOMP can be used to recover all the basis functions whose coefficients are sufficiently large provided that the maximum total correlation between the residual and the remaining basis functions is small.

Besides the SOMP, basis pursuit can also be generalized to obtain a simultaneous approximation version. However, the computational burden is not tractable when facing large scale datasets as in our case. We omit it here, for more details, see Tropp et al. (2006b).

D. Robust basis pursuit using the LAD-Lasso

The previous results assume that the obtained galaxy spectra does not have a heavy tail distribution. However, due to the experiment measurement error, it's very likely that there are outliers in the spectrum. To tackle this problem, more robust methods are needed. For this, we adopt a robust basis pursuit method, named least absolute deviation Lasso (Wang et al., 2004). The basic idea for LAD-Lasso is trying to combine the power of Least absolute deviation regression and the Lasso regression. Instead of using least square as a criterion, LAD-Lasso uses l_1-norm for both the objective function and the penalization constraint. The LAD-Lasso estimator is obtained as

$$\widehat{\beta}^L = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \sum_{j=1}^{p} \beta_j \phi_j(x_i) \right| + \lambda \|\beta\|_1 \tag{4.38}$$

Computationally, the LAD-lasso estimator is quite easy to get. For this, an augmented model approach can be used, i.e. consider an augmented dataset.$Y_i^* \in \mathbb{R}^{n+p}$ and $\mathcal{D}^* =$

$[\mathcal{D}, \phi_{p+1}, ..., \phi_{p+n}]$. where $\phi_{p+j} = n\lambda e_j$ (e_j is the j-th canonical basis function in $\mathbb{R}^n$). Then, it's very easy to see that

$$\widehat{\beta}^L = \arg\min_{\beta} \frac{1}{n+p} \sum_{i=1}^{n} \left| Y_i^* - \sum_{j=1}^{n+p} \beta_j \phi *_j (x_i) \right| \quad \text{for} \quad \phi_j^* \in \mathcal{D}^* \tag{4.39}$$

In fact, the LAD-Lasso can be further generalized into

$$\widehat{\beta}^{LAD} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \sum_{j=1}^{p} \beta_j \phi_j(x_i) \right| + \frac{1}{p} \sum_{k=1}^{p} \lambda_k |\beta_k| \tag{4.40}$$

To obtain $\widehat{\beta}^{LAD}$, the same augmented model approach can be applied. To choose the tuning parameter λ_j, $j = 1, ..., p$. A heuristic suggestion is

$$\widehat{\lambda}_j = \frac{\log(n)}{n \left| \widetilde{\beta}_j \right|} \tag{4.41}$$

where $\widetilde{\beta}_j$ is the ordinary least absolute deviation estimate for the j-th covariant.

The LAD-Lasso doesn't have the well-developed theoretical bounds yet, but some asymptotic properties can be derived under some regularity conditions. Assuming the regression coefficient vector β can be decomposed into two components: $\beta = (\beta_a^T, \beta_b^T)^T$, where $\beta_a = (\beta_1, ..., \beta_{p_0})$ corresponds to the relevant dimensions, $\beta_b = (\beta_{p_0+1}, ..., \beta_p)$ corresponds to the irrelevant ones. Define

$$a_n = \max\{\lambda_j, 1 \le j \le p_0\} \quad \text{and} \quad b_n = \max\{\lambda_j, p_0 < j \le p\} \tag{4.42}$$

The following theorem holds

Theorem 4.7. *Assuming that the independent noise has continuous and positive density at the origin and the covariance matrix of the design matrix exists and is positive definite. Further, if $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$. The LAD-Lasso estimator $\widehat{\beta}^{LAD} = (\widehat{\beta}_a^T, \widehat{\beta}_b^T)^T$ must satisfy*

$$\mathbb{P}\left(\widehat{\beta}_b = 0\right) \longrightarrow 1 \tag{4.43}$$

and

$$\sqrt{n}\left(\widehat{\beta}_a^T - \beta_a^T\right) \longrightarrow \mathcal{N}\left(0, \frac{\Sigma_0^{-1}}{4f^2(0)}\right) \tag{4.44}$$

where Σ_9 is the covariance matrix of all the relevant dimensions and f is the density of the noise.

V. Simultaneous Confidence Band

In the previous sections, we talked about the concrete galaxy spectral function estimation techniques under a unified regression framework. However, for the inference purpose, a more useful thing is the simultaneous confidence bands to cover the true functions $m(x)$. Typically, these bands are of the form

$$\mathcal{I}_x = \left(\widehat{m}(x) + c\sqrt{\mathbf{Var}(\widehat{m}(x))}\ ,\ \widehat{m}(x) + c\sqrt{\mathbf{Var}(\widehat{m}(x))}\right) \tag{5.1}$$

Our sparse composite models are essentially nonparametric, which needs to balance between the bias and the variance. Due to the difficult *bias problem* (Wasserman, 2006), equation (5.1) are in fact confidence bands for $\bar{m}(x) = \mathbb{E}\widehat{m}(x)$, which corresponds to a smoothed version of $m(x)$. Formally, we want to find such a band I such that

$$\mathbb{P}(\bar{m}(x) \in \mathcal{I}_x \text{ for all } x \in X) = \alpha \tag{5.2}$$

Using our sparse composite model approach, since after the basis pursuit, a set of elementary basis functions are selected from an overcomplete dictionary. Further, an ordinary least square regression is applied to estimate the galaxy spectral functions. This corresponds to doing regression in a degenerated linear model. Therefore, all the simultaneous confidence bands developed for the linear models can be applied directly. To ease the following discussions, we redefine a design matrix $\mathbf{X} = [\theta_1, \theta_2, ..., \theta_p]$ as the collection of all the selected basis functions for the degenerated linear model, each θ_i is a length-n vector. We want to find simultaneous confidence bands to cover the smoothed true function $\bar{m}(x) = \mathbb{E}\widehat{Y} = \mathbb{E}\widehat{m}(x)$ with probability α. by the linear model theory, we know that

$$\widehat{m} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) \tag{5.3}$$

Therefore,

$$\bar{m}(x_i) = \mathbb{E}\widehat{m}(x_i) = \theta_i^T\beta \tag{5.4}$$

Now, we have a set $\mathcal{T} = \{\theta_1, ..., \theta_p\}$, and we want to construct a set of confidence intervals such that

$$\mathbb{P}(\text{all } \theta^T\beta \text{ covered, } \theta \in \mathcal{T}) = \alpha \tag{5.5}$$

We call this a "simultaneous set of confidence intervals for $\{\theta^T\beta;\ \ \theta \in \mathcal{T}\}$".

In the remaining part of this section, we will first introduce a nonparametric variance estimation technique, which is especially suitable if the variance is highly heteroscedastic. Then, using the above introduced degenerated linear model framework, four different simultaneous confidence bands will be described.

A. Variance estimation

To estimate simultaneous confidence bands, one important thing is to estimate the standard deviation $\sigma(x_i)$ of at each wavelength value x_i. From figure 1, we have known that the data are highly heteroscedastic. The following approach will be used for variance estimation. Suppose that

$$Y_i = m(x_i) + \sigma(x_i)\epsilon_i, \qquad \epsilon_i \sim^{i.i.d} \mathcal{N}(0,1), \quad i = 1, ..., n \tag{5.6}$$

Let $Z_i = \log\left(Y_i - m(x_i)\right)^2$ and $\delta_i = \log \epsilon_i^2$. Then

$$Z_i = \log\left(\sigma^2(x_i)\right) + \delta_i \tag{5.7}$$

Therefore, we can estimate $\log(\sigma^2(x_i))$ by regressing the log squared residuals on X. For the variance estimation algorithms, there are two places need nonparametric methods to estimate a function. For the first time,since the spectral function is very spiky, ODBP or ODMP are used for the estimation. However, for the residuals, more common methods, like local linear estimator or smoothing splines are more suitable. For more details about this estimation method, see Wasserman (2006).

Variance Estimation: algorithm

1. Estimate $m(x)$ with ODBP/ODMP to get an estimate $\widehat{m}(x)$
2. Define $Z_i = \log\left(Y_i - m(x_i)\right)^2$
3. Regress the Z_i's on the x_i's (using local linear smoother or smoothing splines) to get an estimate $\widehat{q}(x)$ of $\sigma^2(x)$ and let

$$\widehat{\sigma}^2(x) = e^{\widehat{q}(x)} \tag{5.8}$$

Figure 3: The variance function estimation algorithm.

B. Simultaneous confidence bands using Bonferroni correction

The Bonferroni approach can be applied when $\mathcal{T}$ is a finite set, $|\mathcal{T}| = k$. Let I_i to represent the event that $\{\theta_j^T\beta \in \mathcal{I}_{x_i}\}$ and I_i^c to represent the event that $\{\theta_j^T\beta \notin \mathcal{I}_{x_i}\}$. then

$$\begin{aligned} \mathbb{P}(I_1, ..., I_n) &= 1 - \mathbb{P}(I_1^c, ..., I_n^c) && (5.9) \\ &\geq 1 - \sum_{i=1}^{n} \mathbb{P}(I_i^c) && (5.10) \end{aligned}$$

Therefore as long as

$$1 - \alpha \geq \sum_{i=1}^{n} \mathbb{P}(I_j^c) \tag{5.11}$$

the intervals cover simultaneously. One way to achieve this is if each interval individually has probability

$$\alpha' \equiv 1 - \frac{1-\alpha}{n} \tag{5.12}$$

of covering its corresponding value. To do this, use the same approach as used to construct single confidence intervals, but with a threshold α'.

C. Scheffé's simultaneous confidence bands

The Scheffé approach can be applied if $\mathcal{T}$ is a linear subspace of $\mathbb{R}^p$. Begin with the pivotal quantity

$$\frac{\theta^T\widehat{\beta} - \theta^T\beta}{\sqrt{\widehat{\sigma}^2\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta}} \tag{5.13}$$

and postulate that a symmetric interval can be found so that

$$\mathbb{P}\left(-M_\alpha \leq \frac{\theta^T\widehat{\beta} - \theta^T\beta}{\sqrt{\widehat{\sigma}^2\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta}} \leq M_\alpha \text{ for all } \theta \in \mathcal{T}\right) = \alpha \tag{5.14}$$

This would be equivalent to

$$\mathbb{P}\left(\sup_{\theta\in\mathcal{T}} \frac{\left(\theta^T\widehat{\beta} - \theta^T\beta\right)^2}{\widehat{\sigma}^2\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta} \leq M_\alpha^2 \text{ for all } \theta \in \mathcal{T}\right) = \alpha \tag{5.15}$$

Since for $\epsilon = (\epsilon_1, ..., \epsilon_n)$

$$\widehat{\beta} - \beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon \tag{5.16}$$

we have

$$\frac{\left(\theta^T\widehat{\beta} - \theta^T\beta\right)^2}{\widehat{\sigma}^2\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta} = \frac{\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta}{\widehat{\sigma}^2\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta} = \frac{M_\theta^T\epsilon\epsilon^T M_\theta}{\widehat{\sigma}^2 M_\theta^T M_\theta} \tag{5.17}$$

where $M_\theta = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta$. Note that

$$\frac{M_\theta^T\epsilon\epsilon^T M_\theta}{\widehat{\sigma}^2 M_\theta^T M_\theta} = \langle\epsilon, M_\theta/\|M_\theta\|\rangle^2/\widehat{\sigma}^2 \tag{5.18}$$

i.e. it is the squared length of the projection of ϵ onto the line spanned by M_θ (divided by $\widehat{\sigma}^2$). To maximize $\langle \epsilon, M_\theta/\|M_\theta\| \rangle^2$, set $\theta = \mathbf{P}\epsilon$, where $\mathbf{P}$ is the projection matrix onto the linear space

$$\left\{ \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\theta \mid \theta \in \mathcal{T} \right\} = \{M_\theta\} \tag{5.19}$$

Therefore

$$\sup_{\theta \in \mathcal{T}} \langle \epsilon, M_\theta/\|M_\theta\| \rangle^2 / \widehat{\sigma}^2 = \frac{\|\mathbf{P}\epsilon\|^2}{\widehat{\sigma}^2} \tag{5.20}$$

and since $\{M_\theta\}$ is in the column space spanned by $\mathbf{X}$, it follows that $\mathbf{P}\epsilon$ and $\widehat{\sigma}^2$ are independent. Morever,

$$\frac{\|\mathbf{P}\epsilon\|^2}{\widehat{\sigma}^2} \sim \chi^2_q \tag{5.21}$$

where $q = \dim(\mathcal{T})$, and as we know

$$\frac{n-p}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-p} \tag{5.22}$$

Thus

$$\frac{\|\mathbf{P}\epsilon\|^2/q}{\widehat{\sigma}^2} \sim F_{q,n-p} \tag{5.23}$$

Let Q_F be the α quartile of the $F_{q,n-p}$ distribution. Then

$$\mathbb{P}\left(\frac{\theta^T\widehat{\beta} - \theta^T\beta}{\sqrt{\theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta}} \leq \widehat{\sigma}\sqrt{qQ_F} \text{ for all } \theta \right) = \alpha \tag{5.24}$$

so

$$\mathbb{P}\left(\theta^T\widehat{\beta} - \widehat{\sigma}\sqrt{qQ_F V_\theta} \leq \theta^T\beta \leq \theta^T\widehat{\beta} + \widehat{\sigma}\sqrt{qQ_F V_\theta} \text{ for all } \theta \right) = \alpha \tag{5.25}$$

defines a level α simultaneous confidence set for $\{\theta^T\beta | \theta \in \mathcal{T}\}$, where $V_\theta = \theta^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\theta$.

D. Simultaneous confidence bands using the Volume-of-Tube formula

The simultaneous bands using the Volume-of-Tube formula are mainly developed by Sun and Loader (1994). Consider the confidence bands in formula 5.1. Under the degenerated linear model framework, we have

$$\bar{m}(x) = \mathbb{E}\widehat{m}(x) = \sum_{i=1}^{n} l_i(x) m(x_i) \tag{5.26}$$

where $\mathbf{L} = \{l_i(x_j)\}_{i=1,\dots,n}^{j=1,\dots,n}$ is the hat matrix for the design $\mathbf{X}$. In our case, it is

$$\mathbf{L} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{5.27}$$

Also

$$\mathbf{Var}\left(\widehat{m}(x)\right) = \sum_{i=1}^{n} \sigma^2(x_i) l_i^2(x) \tag{5.28}$$

To decide the constant c, first, suppose $\sigma(x)$ is known, we have that

$$\begin{aligned}
\mathbb{P}\left(\bar{m}(x) \notin \mathcal{I}_x \text{ for some } x \in [0,1]\right) &= \mathbb{P}\left(\max_{x\in[0,1]} \frac{\widehat{m}(x) - \bar{m}(x)}{\sqrt{\sum_{i=1}^n \sigma^2(x_i) l_i^2(x)}} > c\right) & (5.29)\\
&= \mathbb{P}\left(\max_{x\in[0,1]} \frac{\sum_{j=1}^n \epsilon_j l_j(x)}{\sqrt{\sum_{i=1}^n \sigma^2(x_i) l_i^2(x)}} > c\right) & (5.30)\\
&= \mathbb{P}\left(\max_{x\in[0,1]} |W(x)| > c\right) & (5.31)
\end{aligned}$$

where $W(x) = \sum_{i=1}^n Z_i T_i(x)$, $Z_i = \epsilon_i/\sigma(x_i) \sim \mathcal{N}(0,1)$ and $T_i(x) = l_i(x)/\sqrt{l_i^2(x)}$. Now, $W(x)$ is a Gaussian process. To find c, we need to calculate the distribution of the maximum of a Gaussian process, this can be done using the following formula, which is also known as "Volume-of-Tube" formula

$$\mathbb{P}\left(\max_{x\in[0,1]} \left|\sum_{i=1}^{n} Z_i T_i(x)\right| > c\right) \approx 2\left(1 - \Phi(x)\right) + \frac{\kappa_0}{\pi} e^{-c^2/2} \tag{5.32}$$

for large c, where

$$\kappa_0 = \int_a^b \|T'(x)\| dx \quad \text{where} \quad T'(x) = \left(\frac{\partial T_1(x)}{\partial x}, \dots, \frac{\partial T_n(x)}{\partial x}\right) \tag{5.33}$$

More details about this formula can be found in Wasserman (2006). If we choose c to solve

$$2\left(1 - \Phi(x)\right) + \frac{\kappa_0}{\pi} e^{-c^2/2} = \alpha \tag{5.34}$$

then the desired simultaneous confidence bands are obtained. If $\sigma(x)$ is unknown, the estimated function $\widehat{\sigma}(x)$ is used. Sun and Loader also suggest to replace the right-hand side of equation 5.32 with

$$\mathbb{P}\left(|T_m| > c\right) + \frac{\kappa_0}{\pi}\left(1 + \frac{c^2}{m}\right)^{-m^2/2} \tag{5.35}$$

where T_m has a t-distribution with $m = n - \text{trace}(\mathbf{L})$ degrees of freedom. For our galaxy spectra analysis task, n is fairly large (≈ 4000), equation 5.32 is already a suitable approximation.

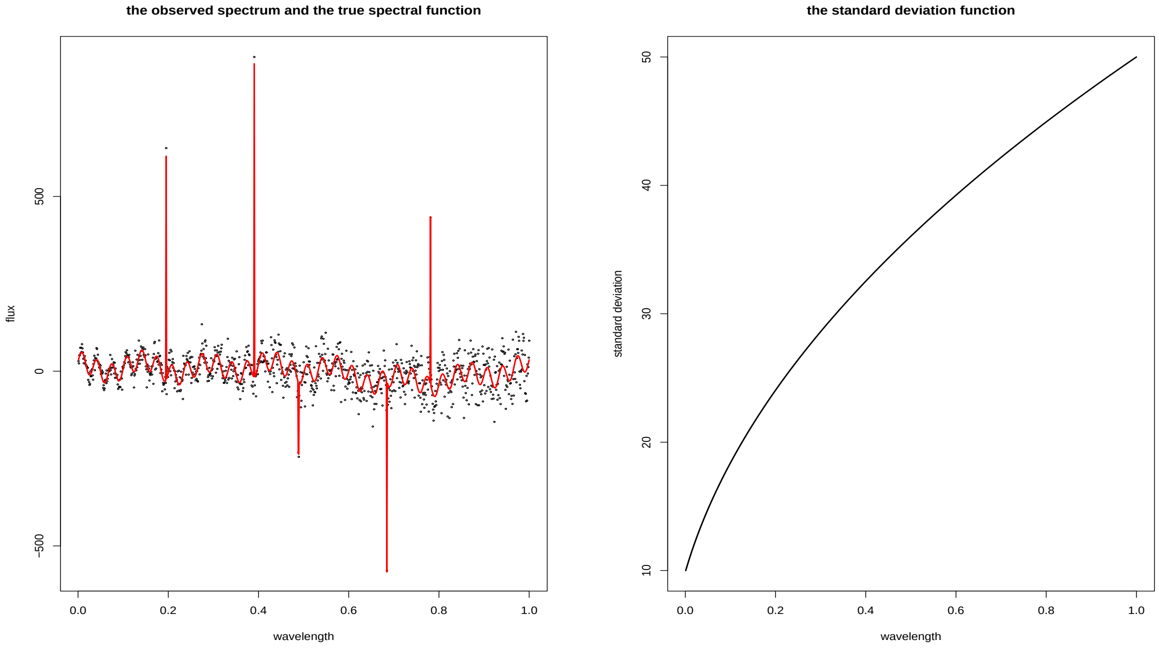
E. Simultaneous confidence bands using bootstrapping

After we fit the galaxy spectral function , $\widehat{m}(x)$, the residuals $\widehat{\epsilon}_i = Y_i - \widehat{m}(x_i)$ can be calculated. Resample the residuals to get bootstrap residuals $\widehat{\epsilon}_1^*, \widehat{\epsilon}_2^*, ..., \widehat{\epsilon}_n^*$. Now let $Y_i^* = \widehat{m}(x_i) + \epsilon_i^*$, $i = 1, ..., n$. We re-estimate the spectral function, repeat this process B times, and get the upper and lower $\alpha/2$ quartiles at each point to calculate the simultaneous confidence band. This approach is very computationally intensive, but is still affordable.

VI. Experimental Results

Different methods are tested on both synthetic and real-wold datasets, and has been proved to be both effective and easy to implement. In this section, we first use some synthetic data to illustrate the performance of different function and confidence bands estimators. Then, the real-world data from SDSS is ued to test the performance of different methods.

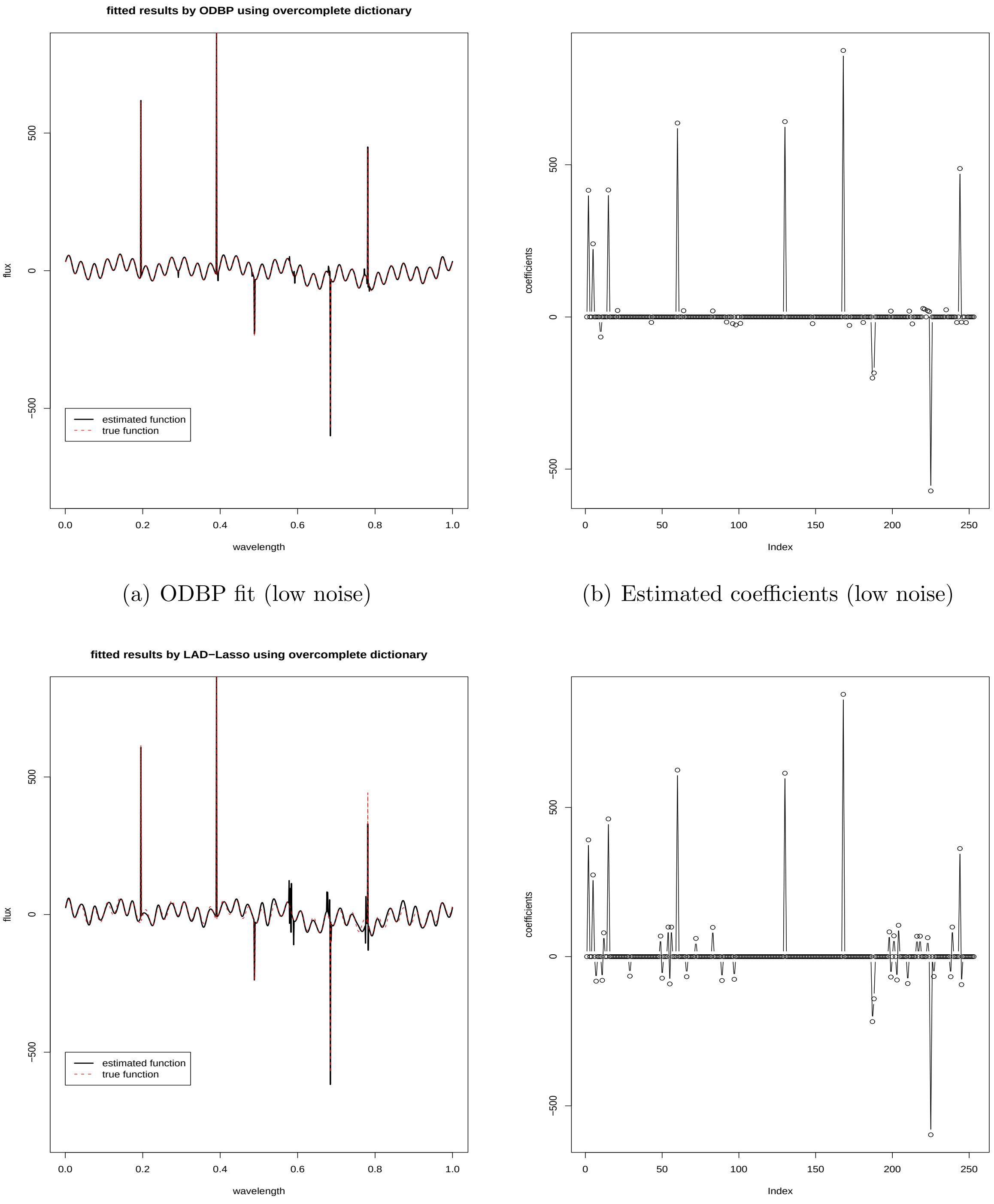
A. Synthetic dataset



(a) the synthetic spectrum

(b) standard deviation function

Figure 4: *Left: the synthetic spectrum using the Fourier basis and the narrow Gaussian density functions, Right: the heteroscedastic variance function*



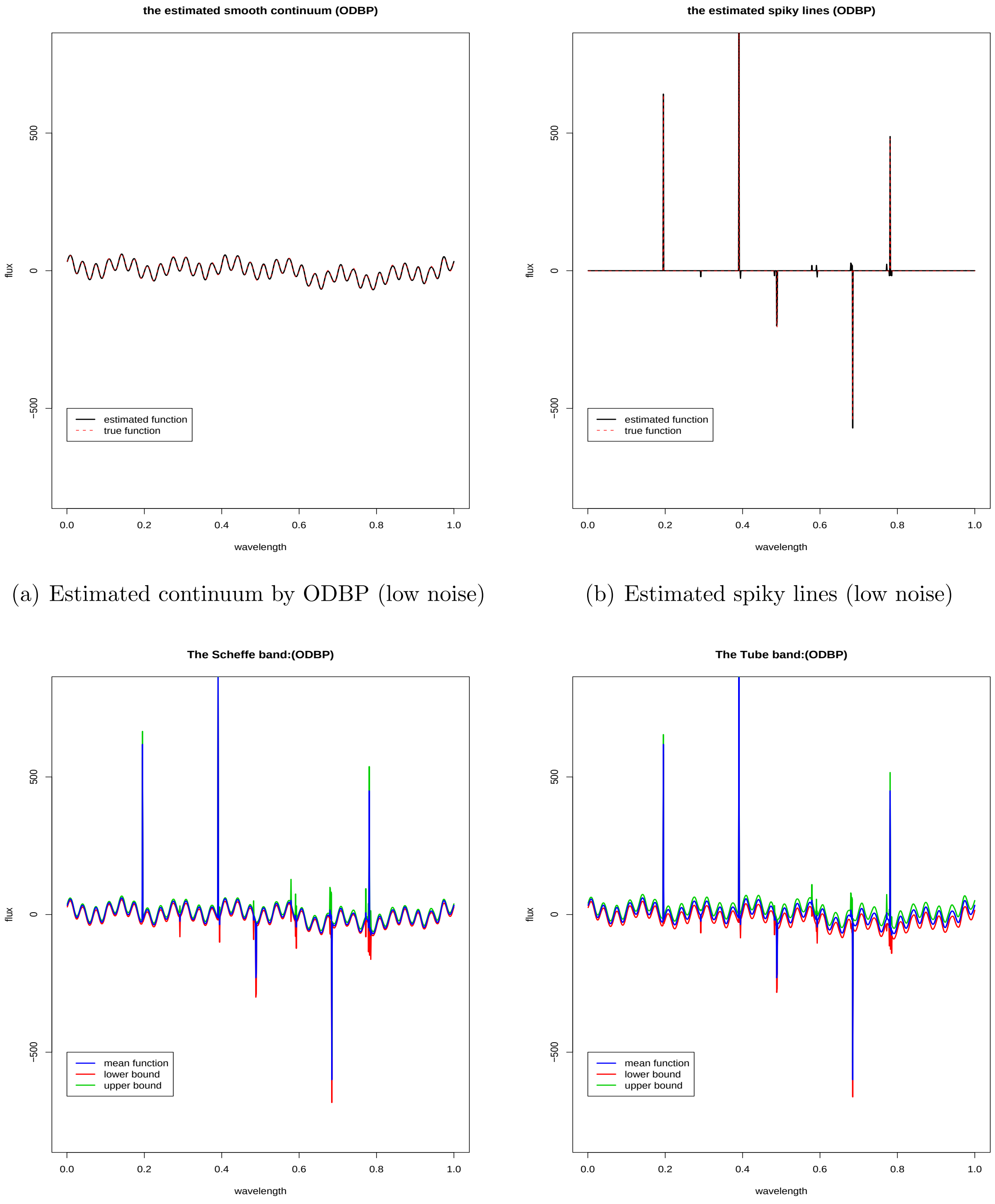
(a) ODBP fit (low noise)

(b) Estimated coefficients (low noise)

(c) ODBP fit (high noise)

(d) Estimated coefficients (high noise)

Figure 5: *Upper: the estimated spectral function and the corresponding coefficients plot from the ODBP for the dataset with a low noise level. Lower: the same results from the ODBP for the dataset with a high noise level*



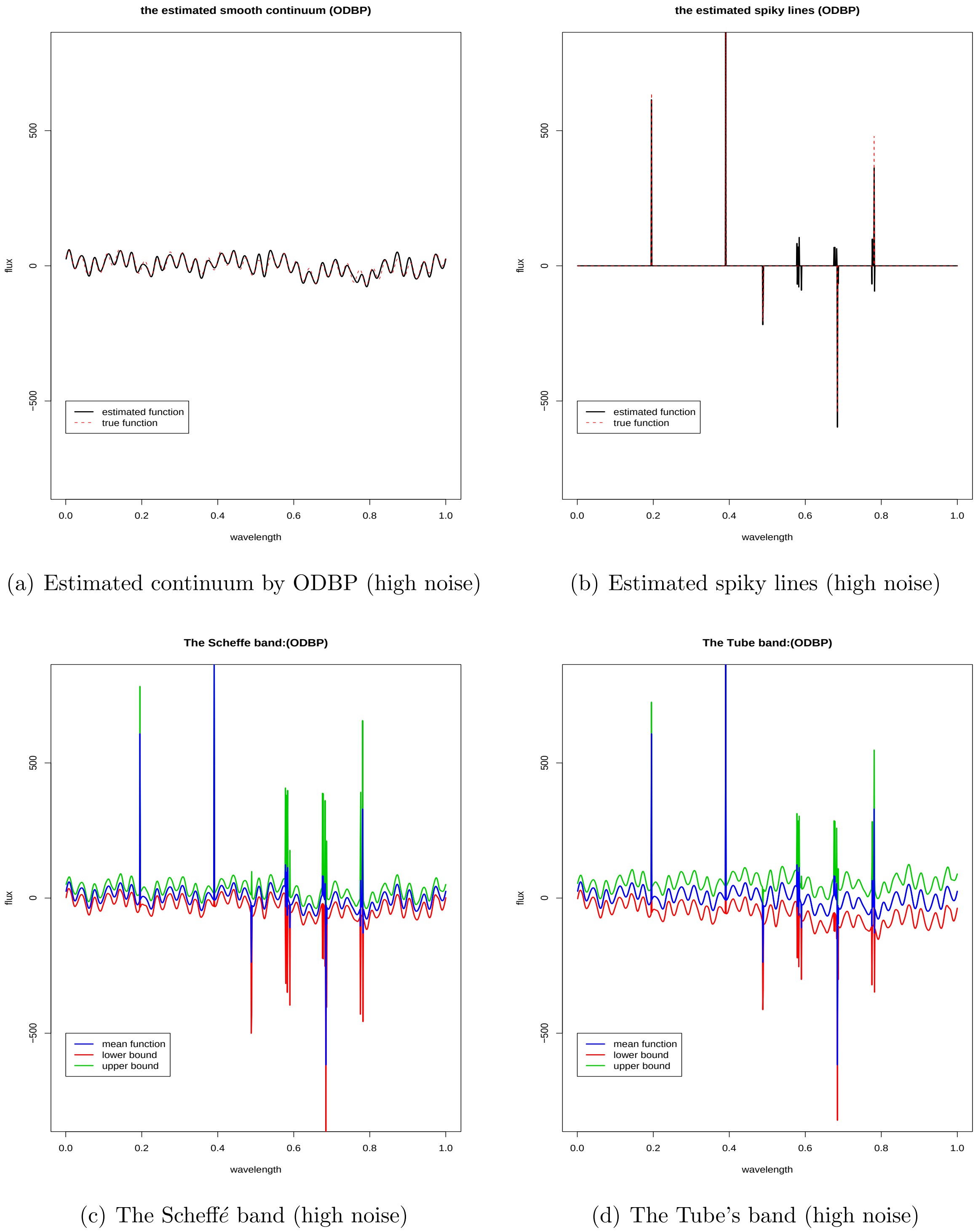
(a) Estimated continuum by ODBP (low noise)

(b) Estimated spiky lines (low noise)

(c) The Scheffé band (low noise)

(d) The Tube's band (low noise)

Figure 6: *Upper: the estimated smooth continuum and the spiky lines from the ODBP for the dataset with a low noise level. Lower: simultaneous confidence bands from the ODBP for the dataset with a low noise level*



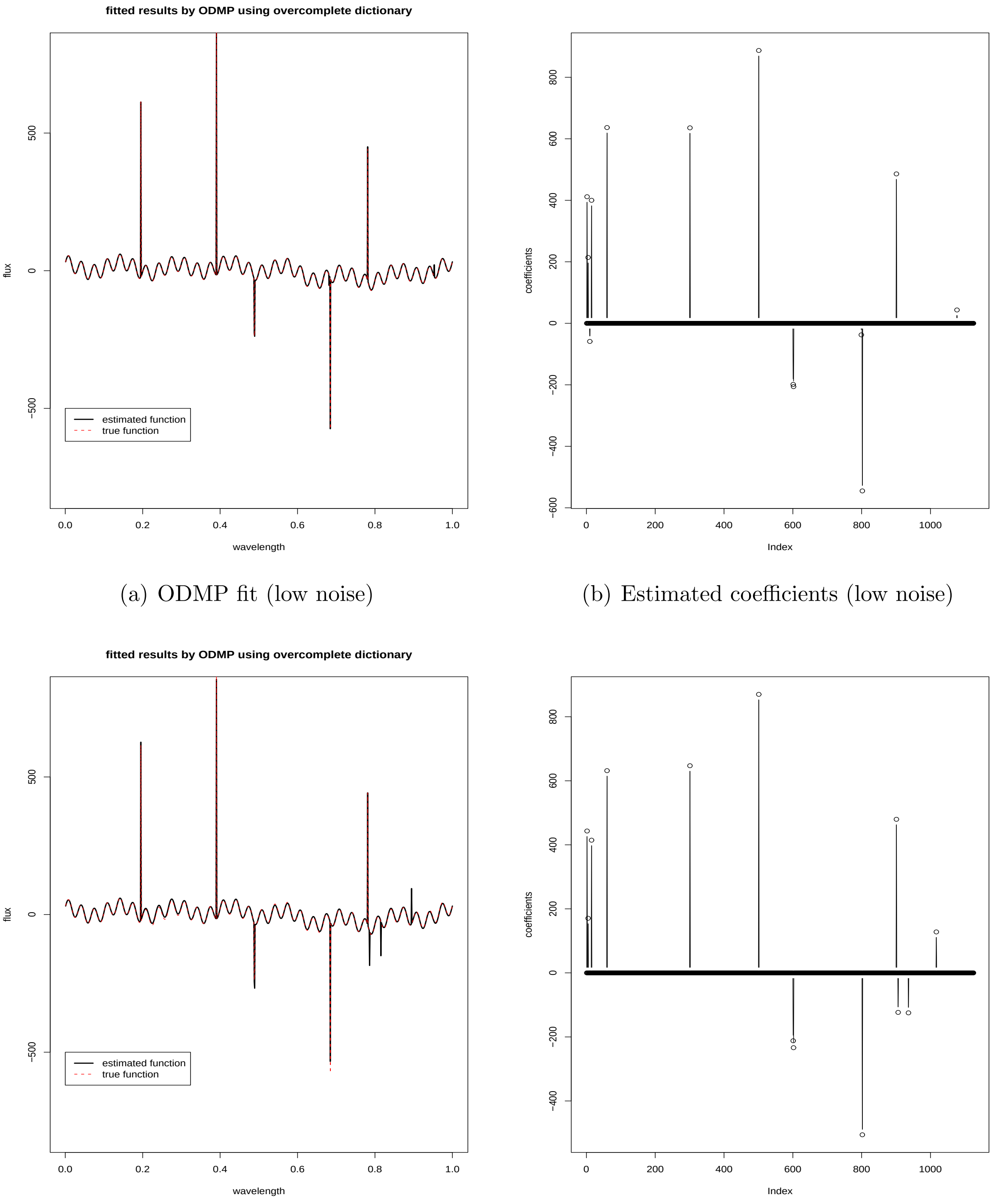
(a) Estimated continuum by ODBP (high noise)

(b) Estimated spiky lines (high noise)

(c) The Scheffé band (high noise)

(d) The Tube's band (high noise)

Figure 7: *Upper: the estimated smooth continuum and the spiky lines from the ODBP for the dataset with a high noise level. Lower: simultaneous confidence bands from the ODBP for the dataset with a high noise level*



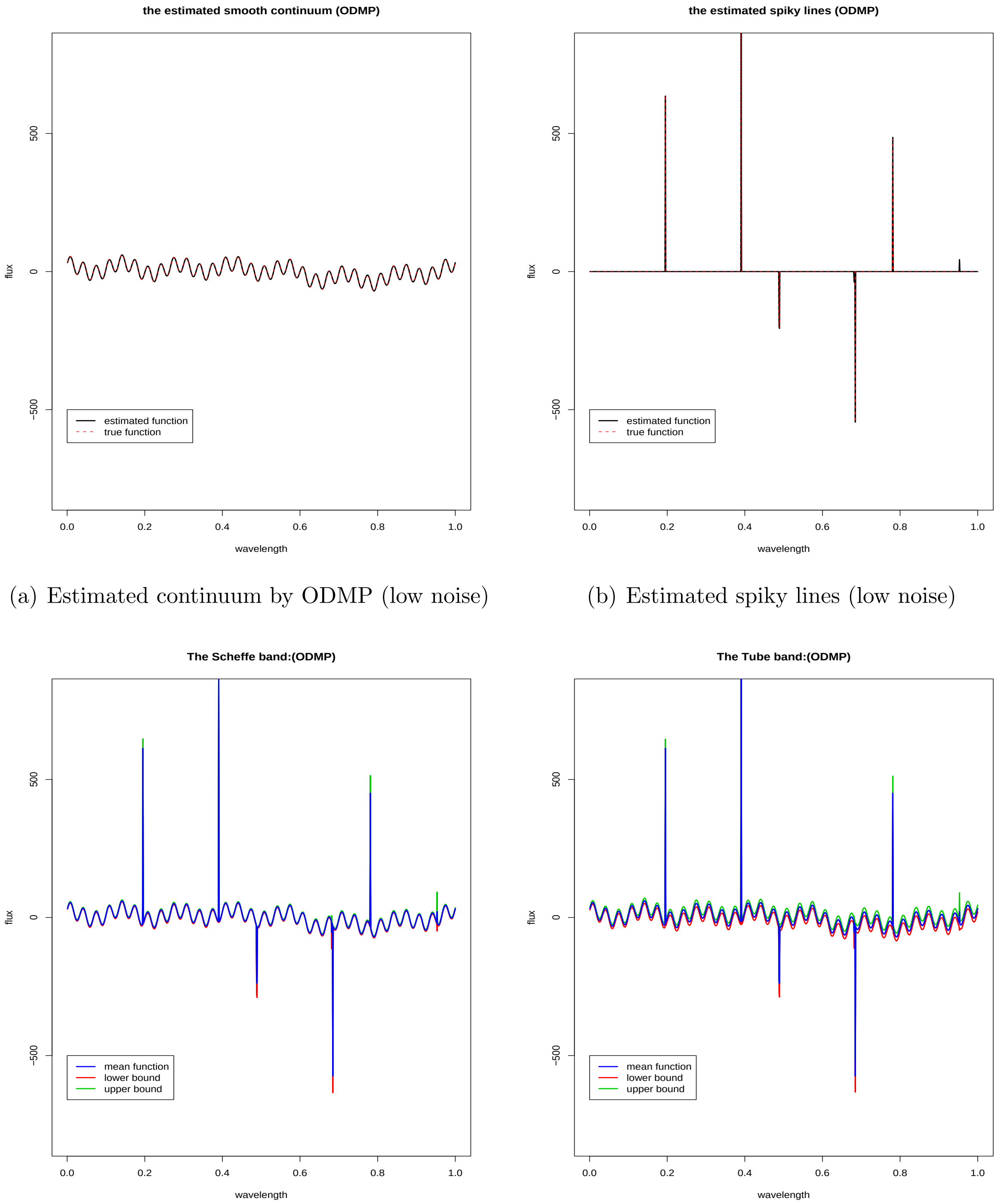
(a) ODMP fit (low noise)

(b) Estimated coefficients (low noise)

(c) ODMP fit (high noise)

(d) Estimated coefficients (high noise)

Figure 8: *Upper: the estimated spectral function and the corresponding coefficients plot from the ODMP for the dataset with a low noise level. Lower: the same results from the ODMP for the dataset with a high noise level*



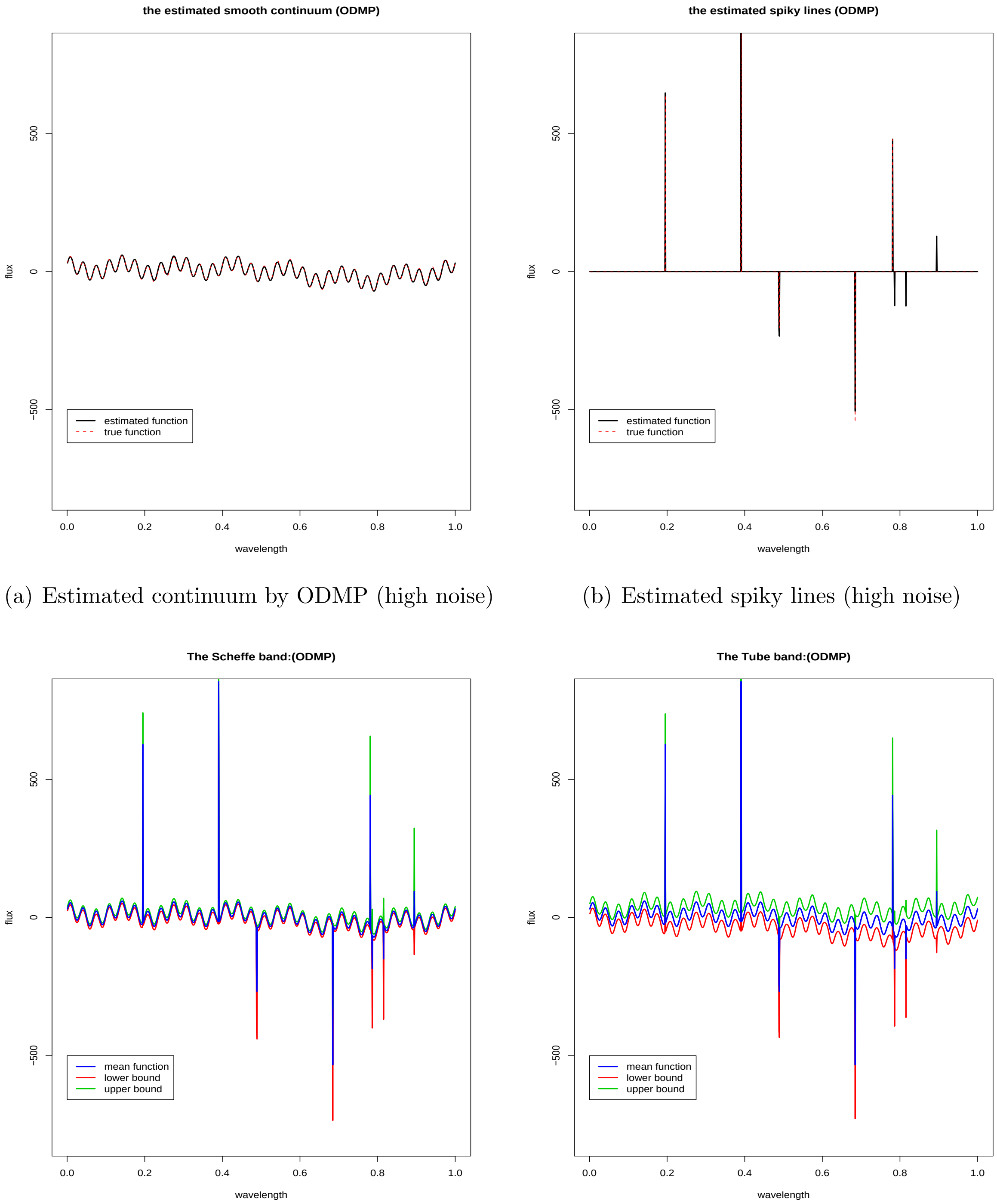
(a) Estimated continuum by ODMP (low noise)

(b) Estimated spiky lines (low noise)

(c) The Scheffé band (low noise)

(d) The Tube's band (low noise)

Figure 9: *Upper: the estimated smooth continuum and the spiky lines from the ODMP for the dataset with a low noise level. Lower: simultaneous confidence bands from the ODMP for the dataset with a low noise level*



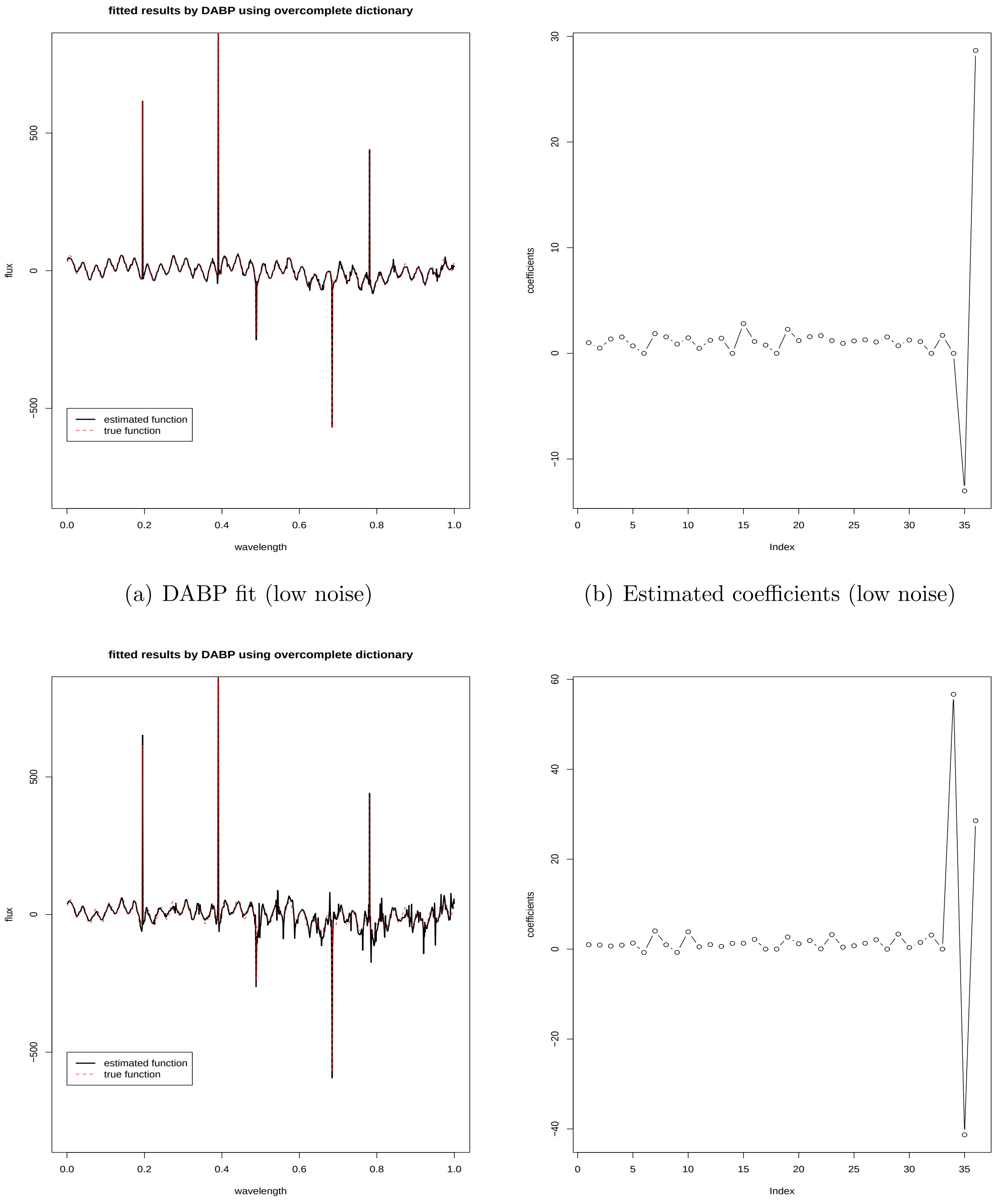
(a) Estimated continuum by ODMP (high noise)

(b) Estimated spiky lines (high noise)

(c) The Scheffé band (high noise)

(d) The Tube's band (high noise)

Figure 10: *Upper: the estimated smooth continuum and the spiky lines from the ODMP for the dataset with a high noise level. Lower: simultaneous confidence bands from the ODMP for the dataset with a high noise level*



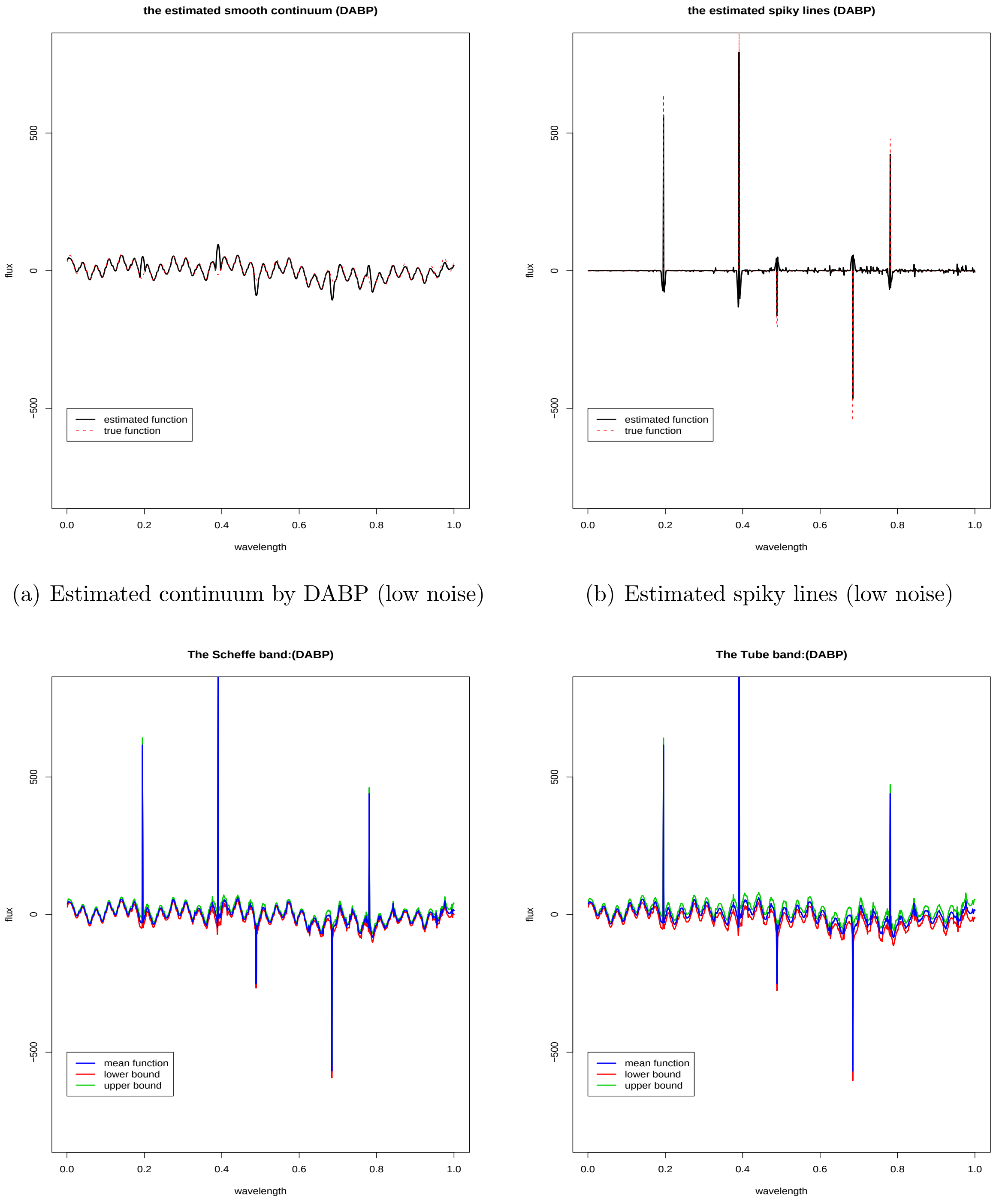
(a) DABP fit (low noise)

(b) Estimated coefficients (low noise)

(c) DABP fit (high noise)

(d) Estimated coefficients (high noise)

Figure 11: *Upper: the estimated spectral function and the corresponding coefficients plot from the DABP for the dataset with a low noise level. Lower: the same results from the DABP for the dataset with a high noise level*



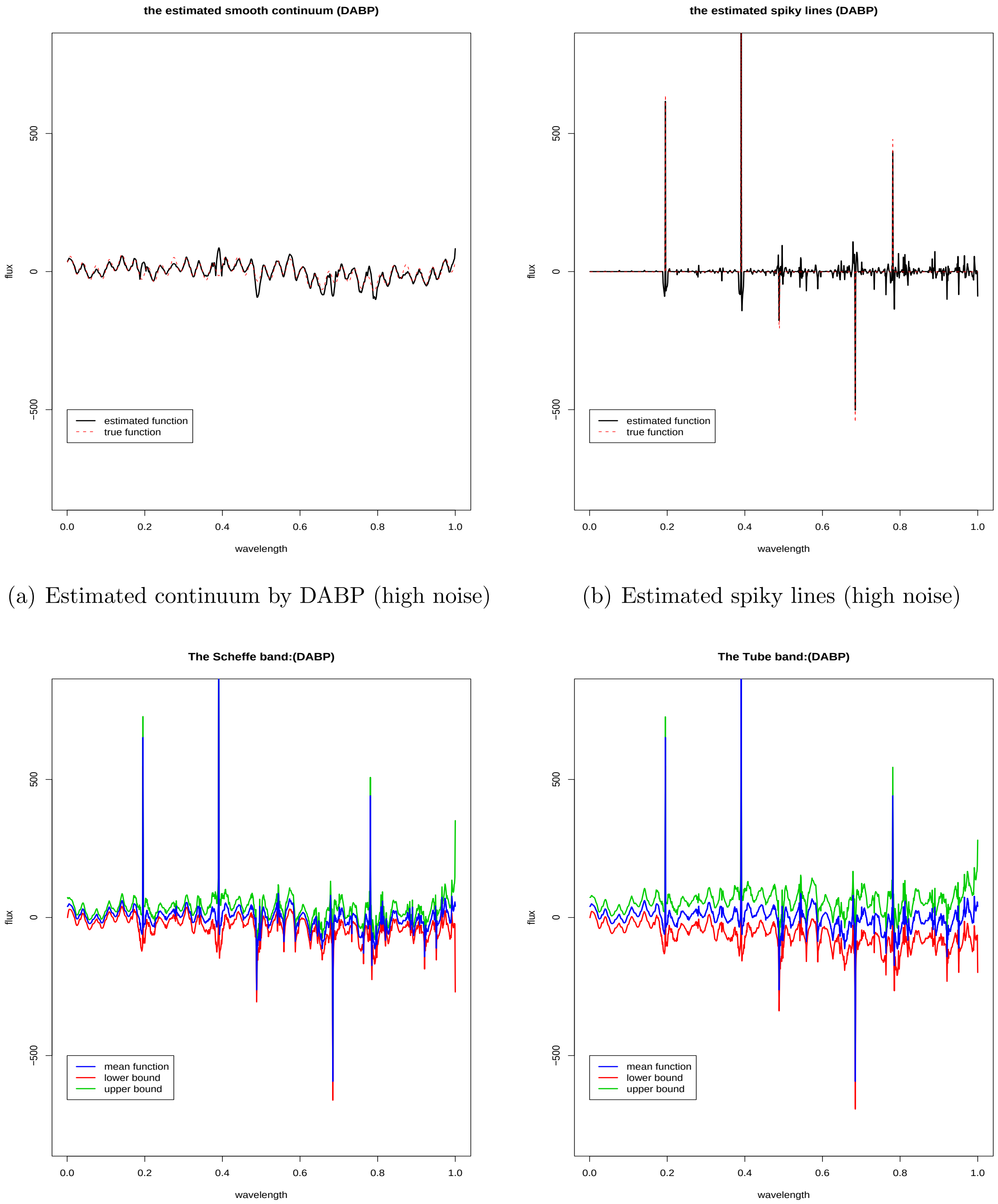
(a) Estimated continuum by DABP (low noise)

(b) Estimated spiky lines (low noise)

(c) The Scheffé band (low noise)

(d) The Tube's band (low noise)

Figure 12: *Upper: the estimated smooth continuum and the spiky lines from the DABP for the dataset with a low noise level. Lower: simultaneous confidence bands from the DABP for the dataset with a low noise level*



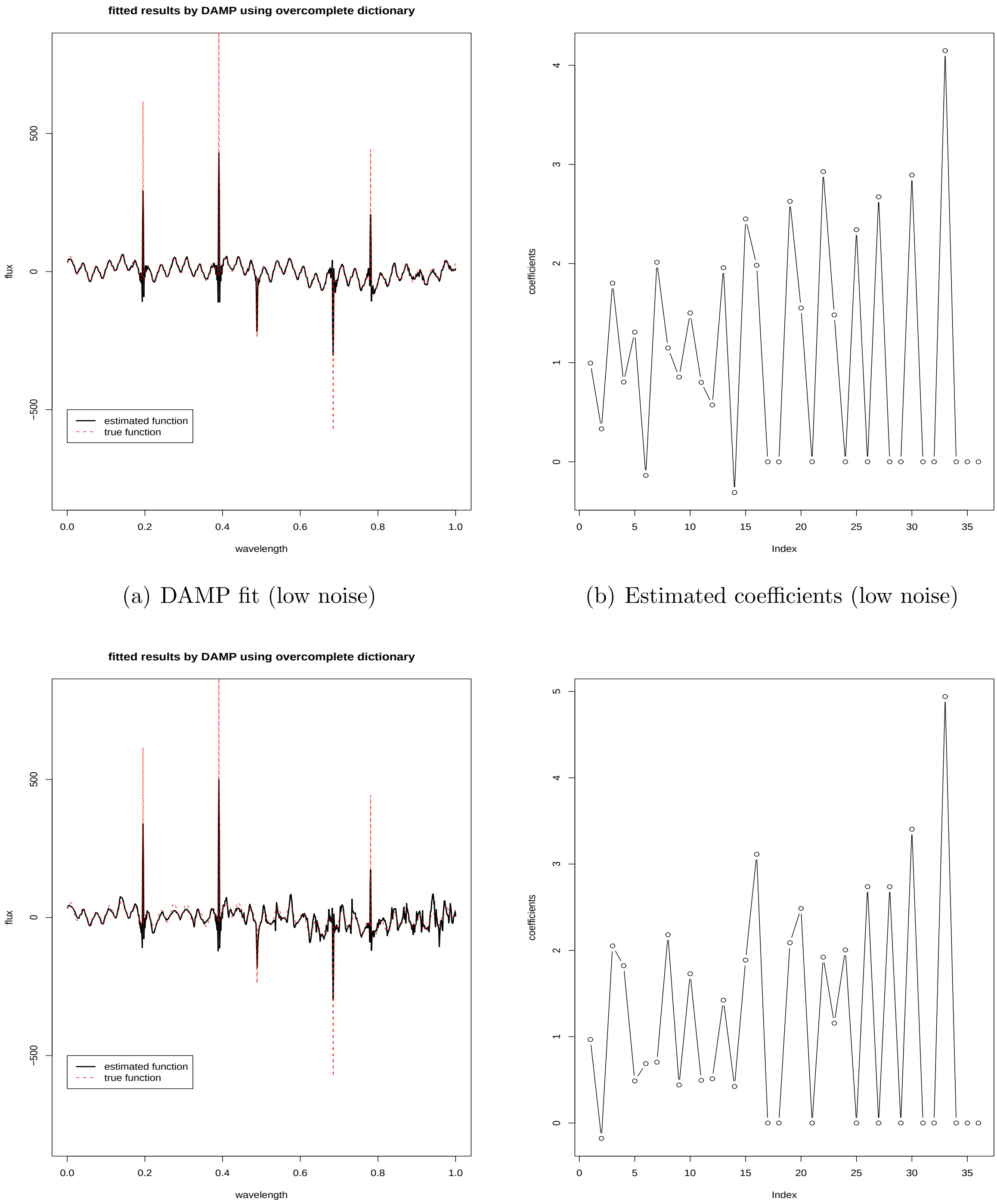
(a) Estimated continuum by DABP (high noise)

(b) Estimated spiky lines (high noise)

(c) The Scheffé band (high noise)

(d) The Tube's band (high noise)

Figure 13: *Upper: the estimated smooth continuum and the spiky lines from the DABP for the dataset with a high noise level. Lower: simultaneous confidence bands from the DABP for the dataset with a high noise level*



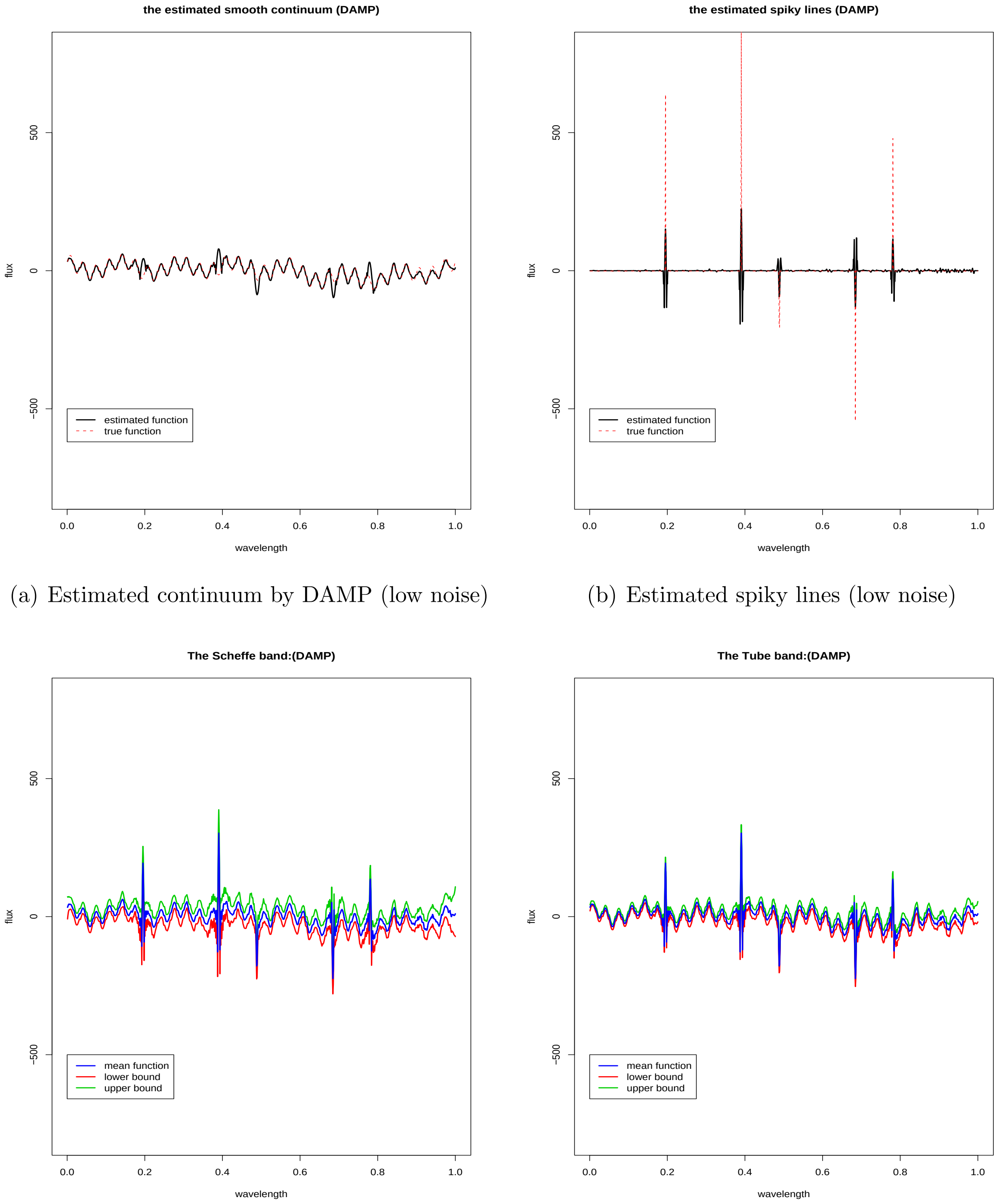
(a) DAMP fit (low noise)

(b) Estimated coefficients (low noise)

(c) DAMP fit (high noise)

(d) Estimated coefficients (high noise)

Figure 14: *Upper: the estimated spectral function and the corresponding coefficients plot from the DAMP for the dataset with a low noise level. Lower: the same results from the DAMP for the dataset with a high noise level*



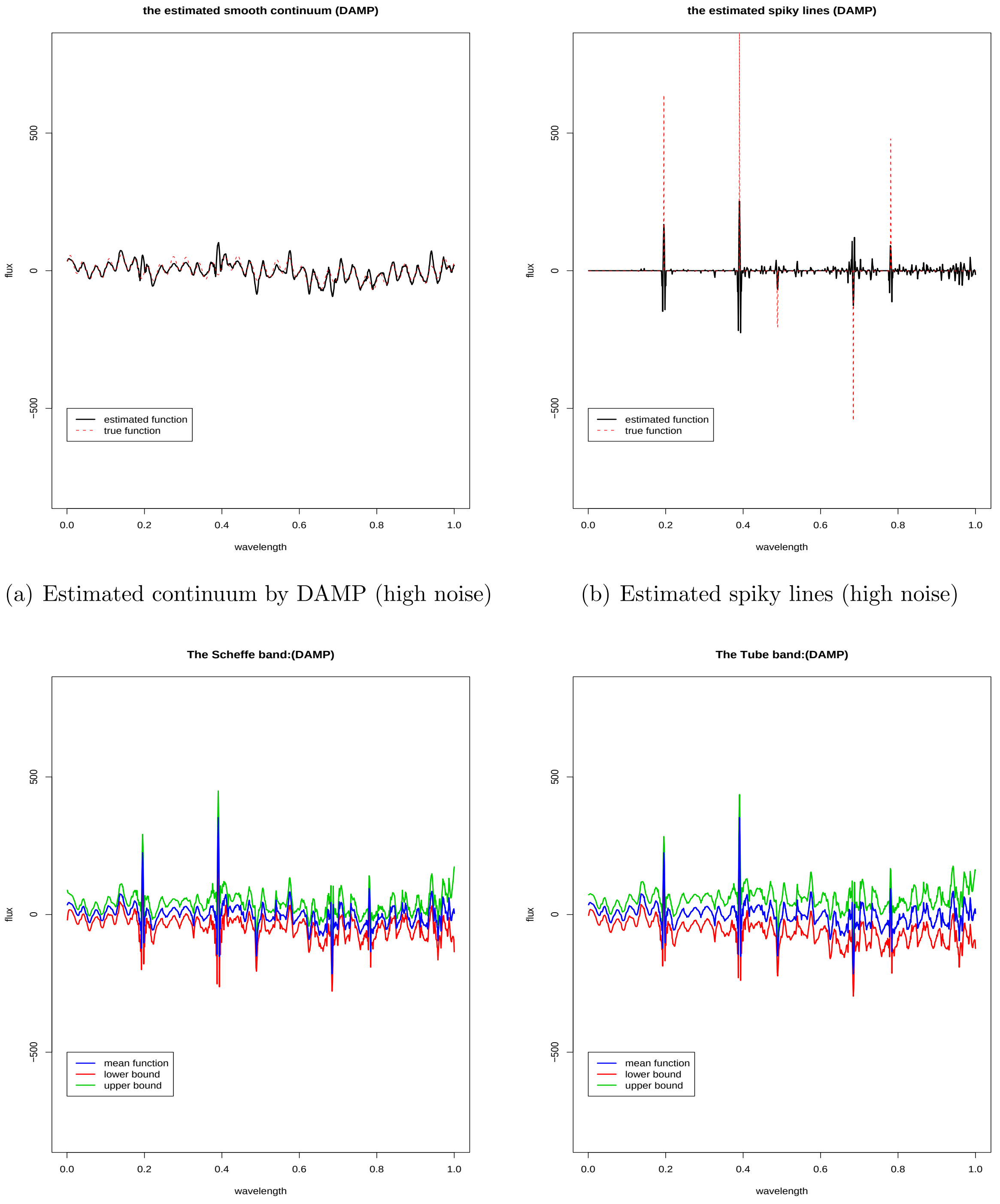
(a) Estimated continuum by DAMP (low noise)

(b) Estimated spiky lines (low noise)

(c) The Scheffé band (low noise)

(d) The Tube's band (low noise)

Figure 15: *Upper: the estimated smooth continuum and the spiky lines from the DAMP for the dataset with a low noise level. Lower: simultaneous confidence bands from the DAMP for the dataset with a low noise level*



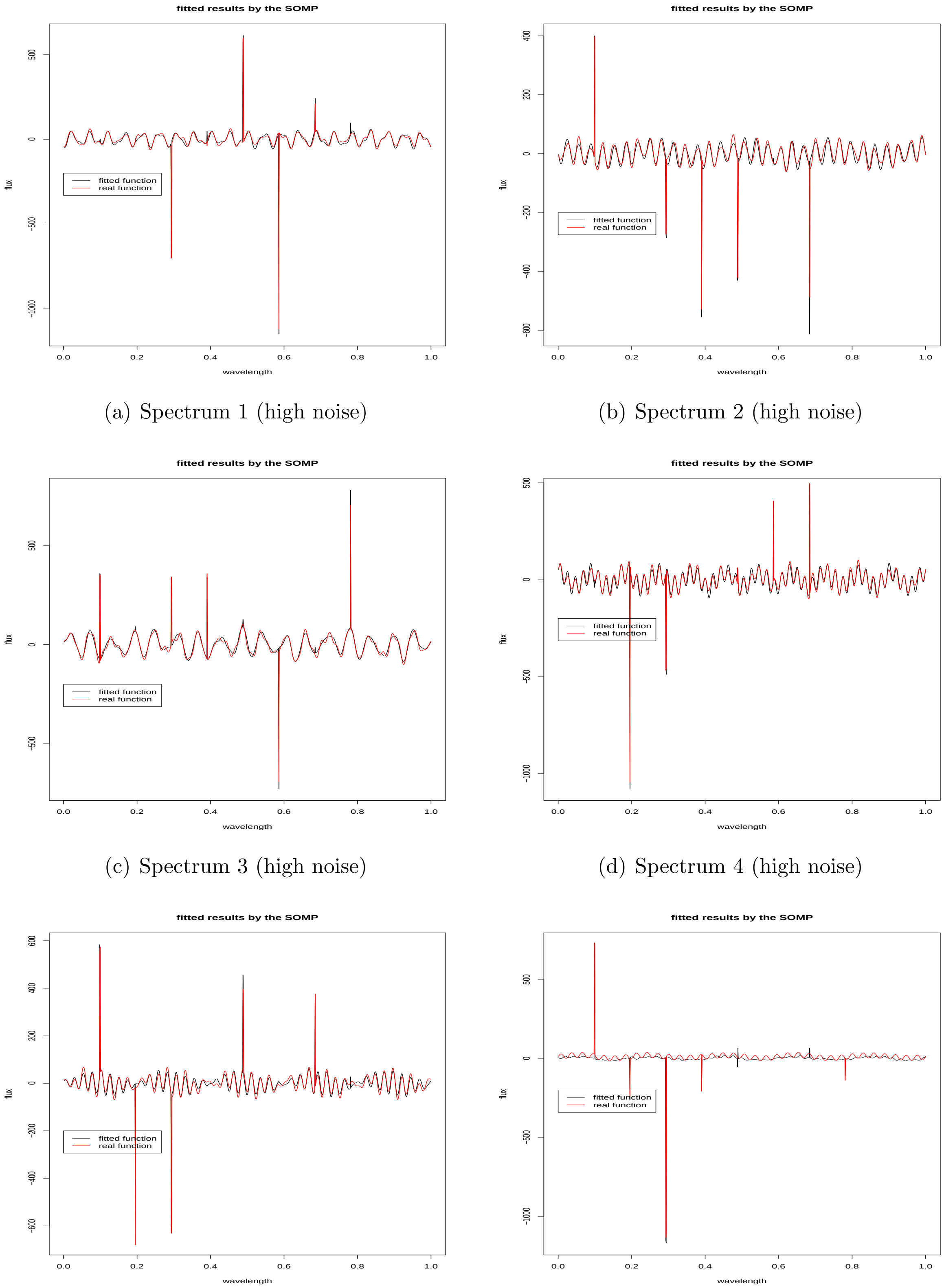
(a) Estimated continuum by DAMP (high noise)

(b) Estimated spiky lines (high noise)

(c) The Scheffé band (high noise)

(d) The Tube's band (high noise)

Figure 16: *Upper: the estimated smooth continuum and the spiky lines from the DAMP for the dataset with a high noise level. Lower: simultaneous confidence bands from the DAMP for the dataset with a high noise level*



(a) Spectrum 1 (high noise)

(b) Spectrum 2 (high noise)

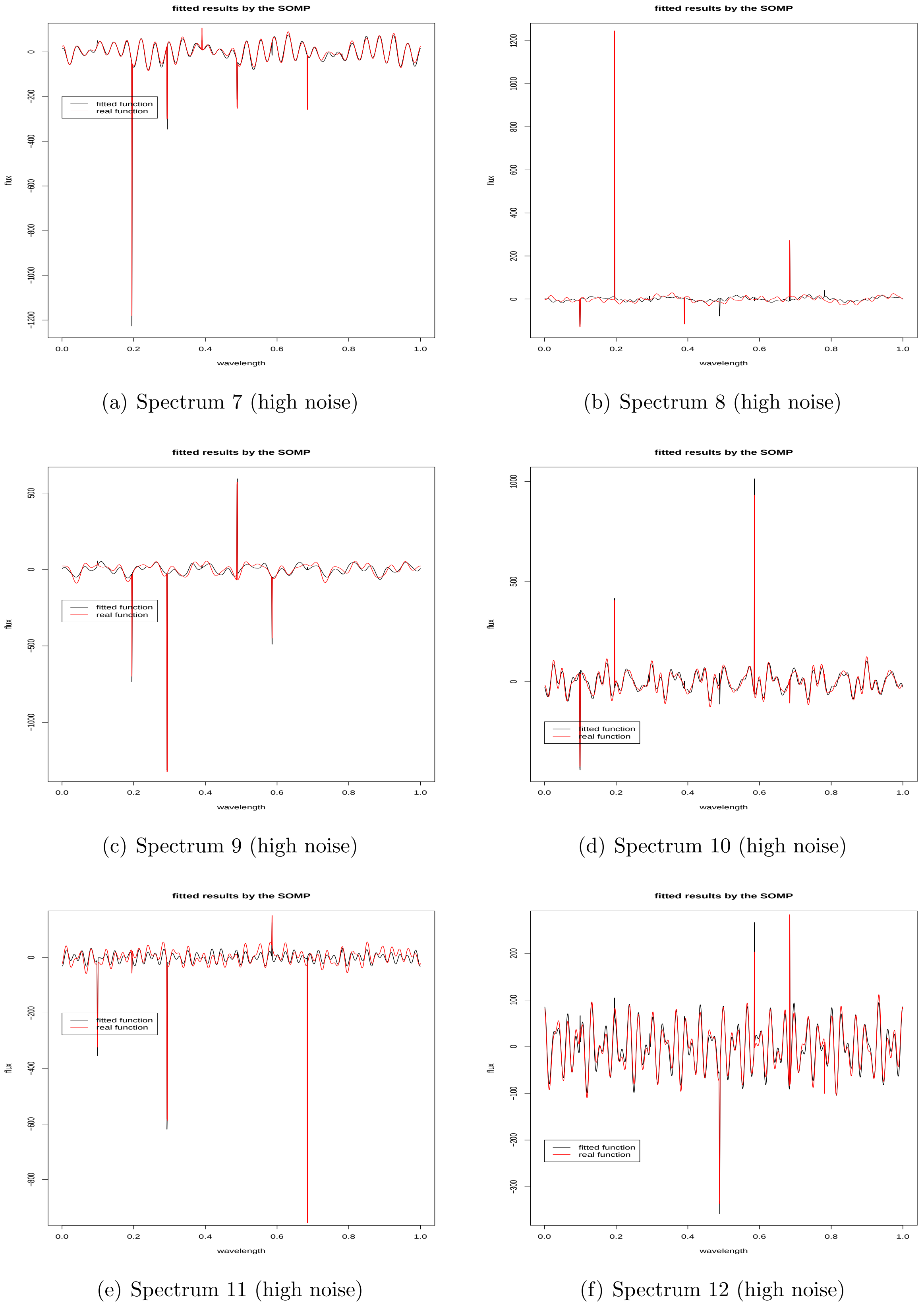
(c) Spectrum 3 (high noise)

(d) Spectrum 4 (high noise)

(e) Spectrum 5 (high noise)

(f) Spectrum 6 (high noise)

Figure 17: *Fitted results for 12 galaxy spectra (Here shows the results for spectra 1 - 6)using the SOMP algorithms. Here, the 12 spectra have a high noise level*



(a) Spectrum 7 (high noise)

(b) Spectrum 8 (high noise)

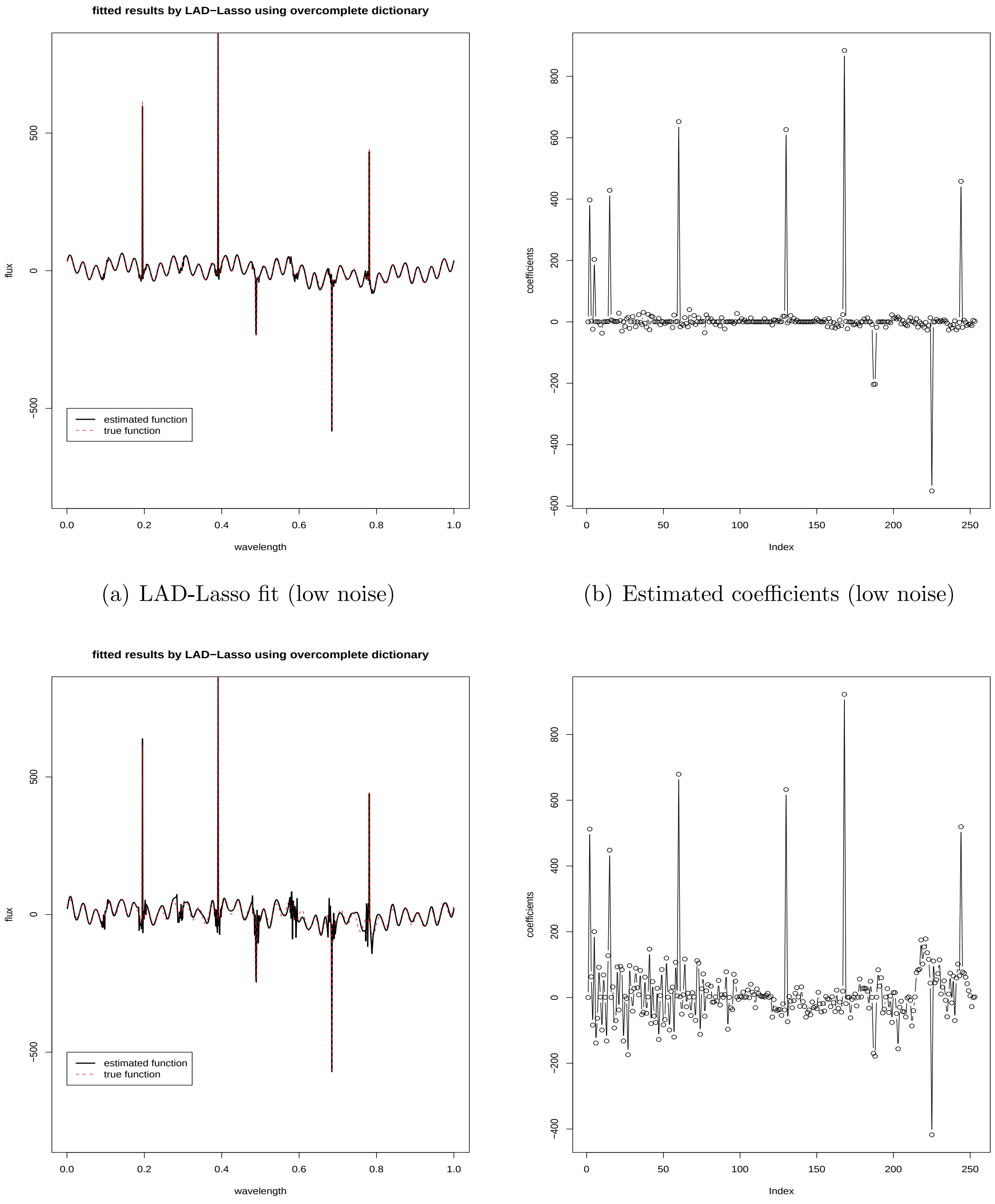
(c) Spectrum 9 (high noise)

(d) Spectrum 10 (high noise)

(e) Spectrum 11 (high noise)

(f) Spectrum 12 (high noise)

Figure 18: *Fitted results for 12 galaxy spectra (Here shows the results for spectra 7 - 12)using the SOMP algorithms. Here, the 12 spectra have a high noise level*



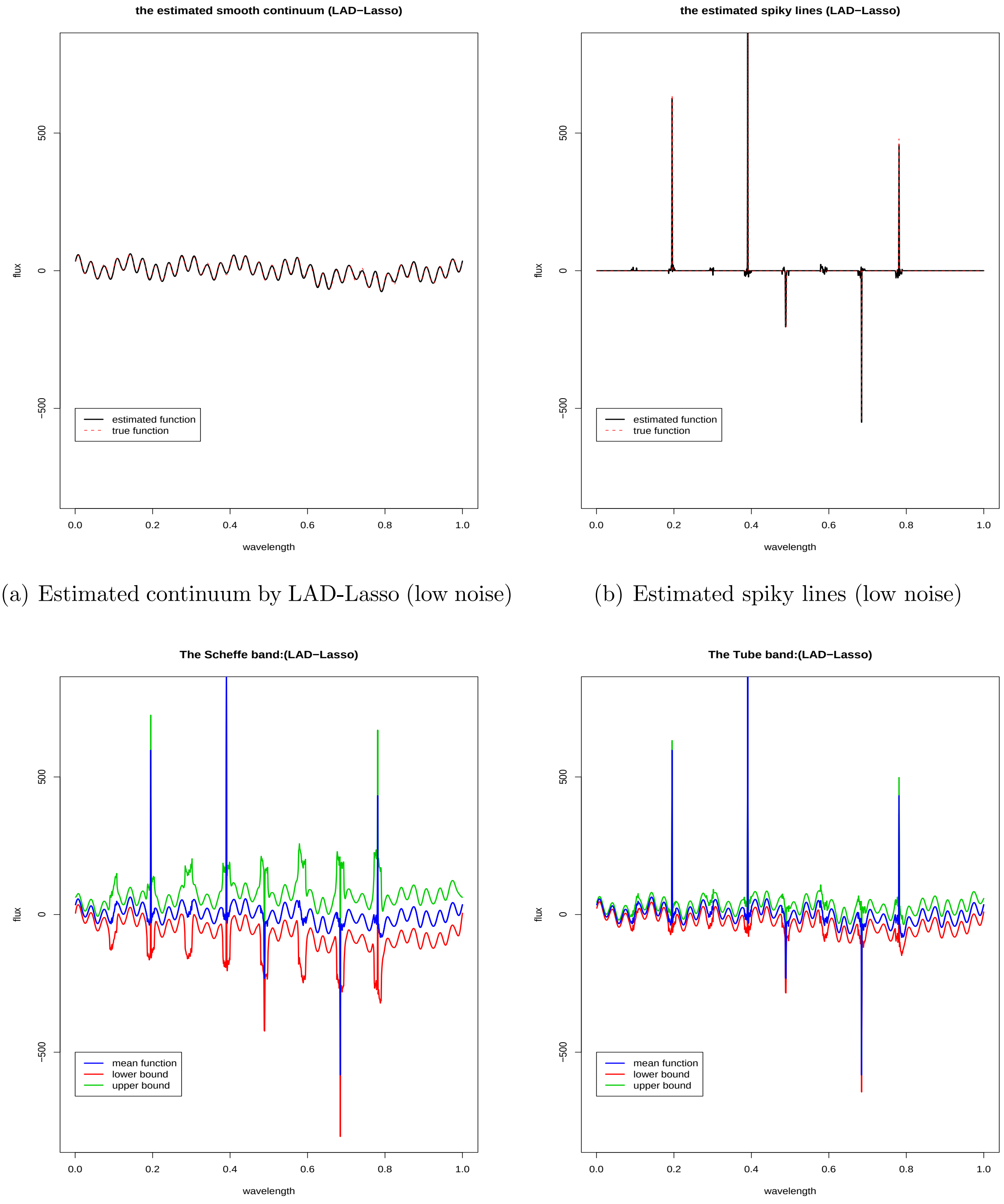
(a) LAD-Lasso fit (low noise)

(b) Estimated coefficients (low noise)

(c) LAD-Lasso fit (high noise)

(d) Estimated coefficients (high noise)

Figure 19: *Upper: the estimated spectral function and the corresponding coefficients plot from the LAD-Lasso for the dataset with a low noise level. Lower: the same results from the LAD-Lasso for the dataset with a high noise level*



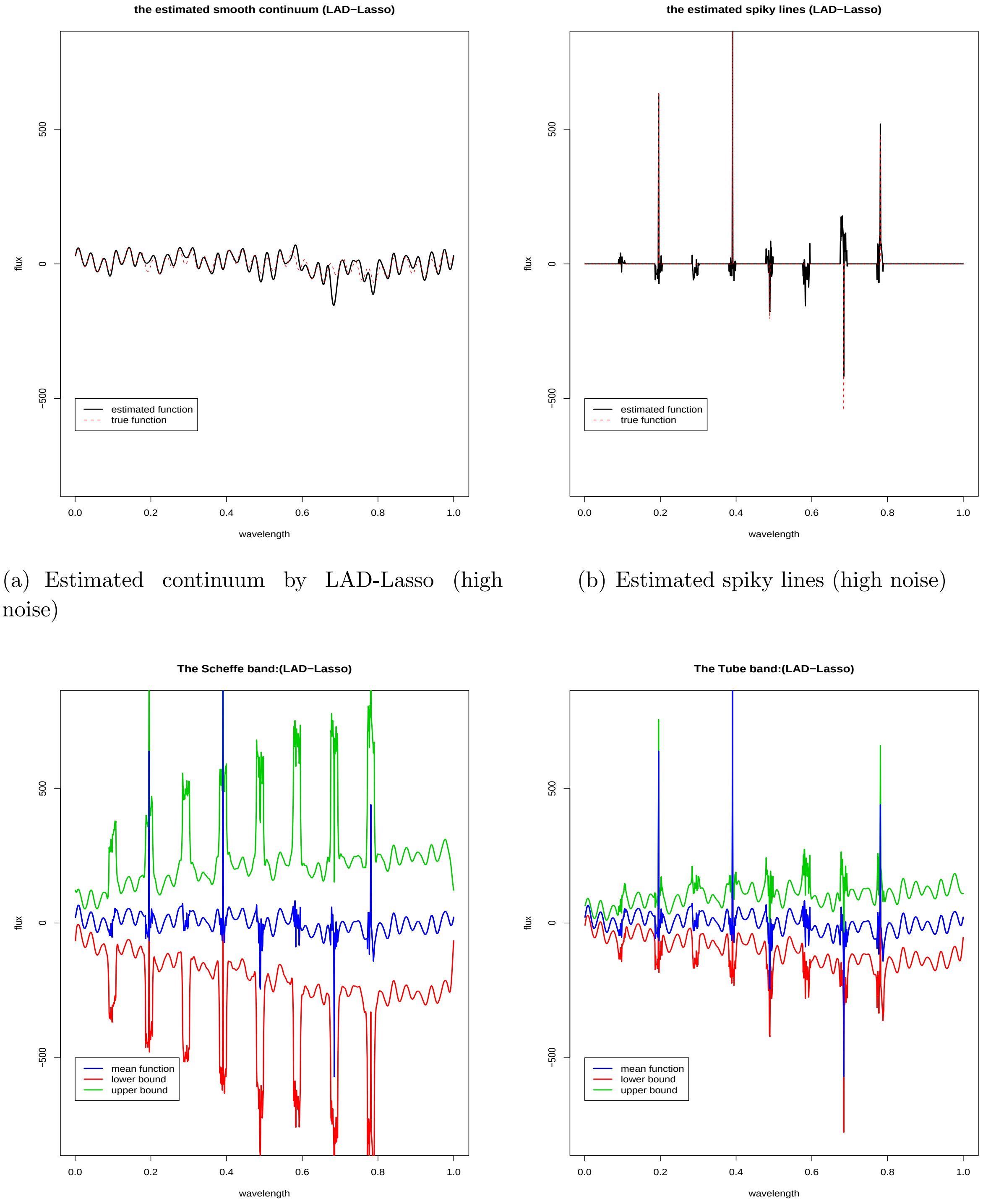
(a) Estimated continuum by LAD-Lasso (low noise)

(b) Estimated spiky lines (low noise)

(c) The Scheffé band (low noise)

(d) The Tube's band (low noise)

Figure 20: *Upper: the estimated smooth continuum and the spiky lines from the LAD-Lasso for the dataset with a low noise level. Lower: simultaneous confidence bands from the LAD-Lasso for the dataset with a low noise level*



(a) Estimated continuum by LAD-Lasso (high noise)

(b) Estimated spiky lines (high noise)

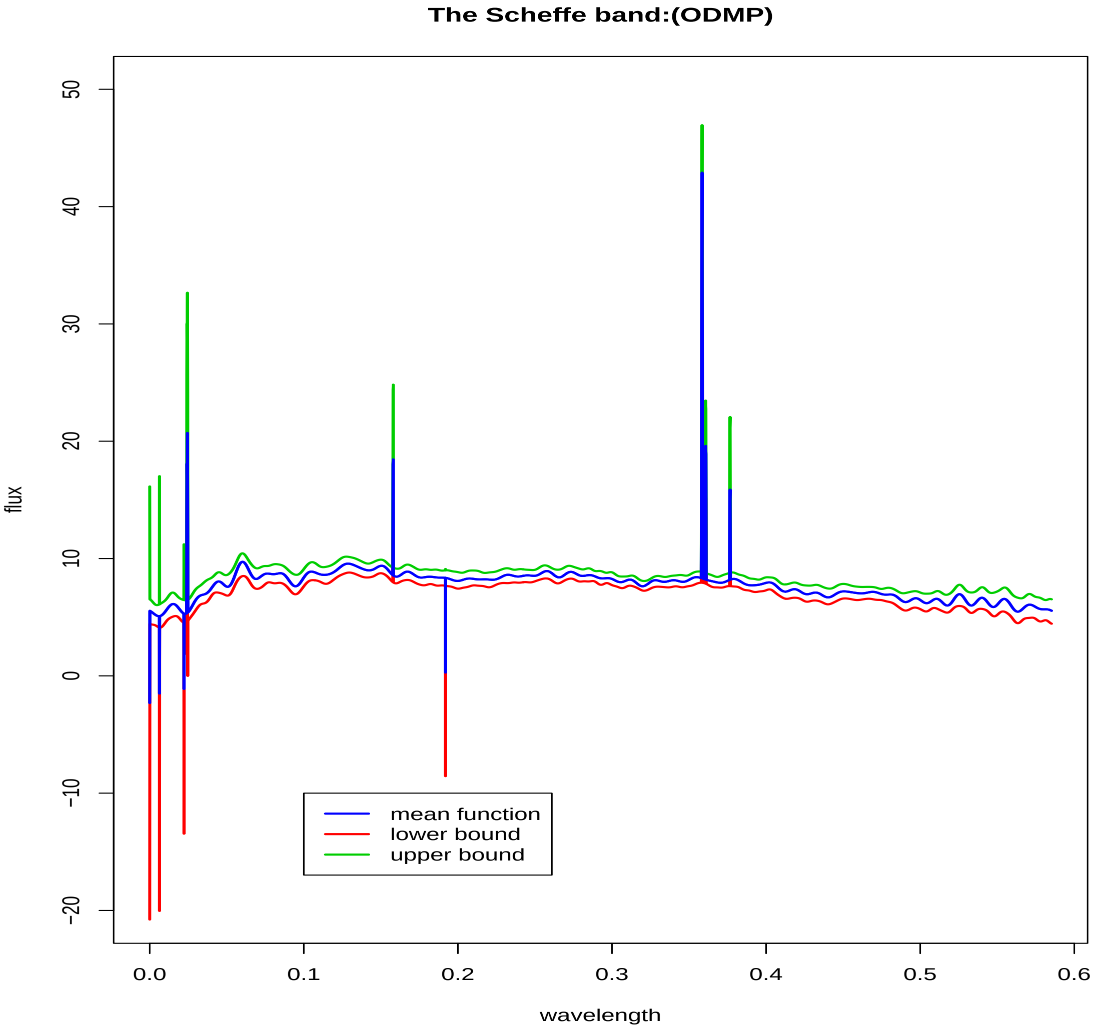
(c) The Scheffé band (high noise)

(d) The Tube's band (high noise)

Figure 21: *Upper: the estimated smooth continuum and the spiky lines from the LAD-Lasso for the dataset with a high noise level. Lower: simultaneous confidence bands from the LAD-Lasso for the dataset with a high noise level*

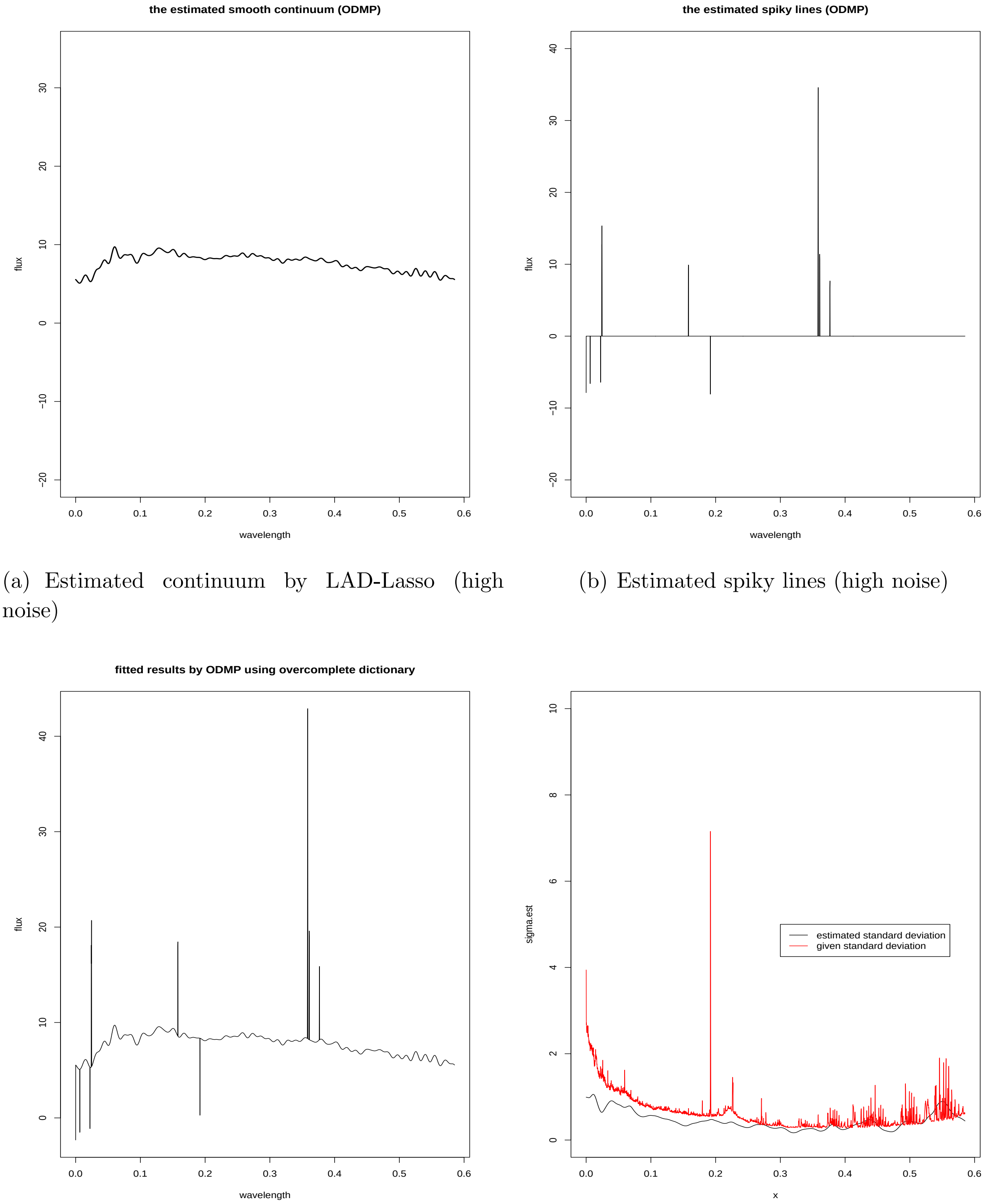
B. Real SDSS Dataset and exploratory data analysis

From the above simulations on the synthetic datasets, the matching pursuit method's performance is better than the basis pursuit approach. Especially, ODMP obtains the best results in both high noise and the low noise cases. In the following, we will apply ODMP to fit the real galaxy spectra. The observed spectra is shown in figure 1. We apply ODMP method using an overcomplete dictionary with Fourier basis functions and the Dirac delta functions. Our iteration number $T = 50$. After the pursuit, we get the fitted smooth continuum and the spiky lines as is shown in figure 22



(a) Estimated continuum by LAD-Lasso (high noise)

Figure 22: *The simultaneous confidence bands (Scheffé) and the fitted galaxy spectral function using the ODMP method*



(a) Estimated continuum by LAD-Lasso (high noise)

(b) Estimated spiky lines (high noise)

(c) The Scheffé band (high noise)

(d) The Tube's band (high noise)

Figure 23: *Upper: the estimated smooth continuum and the spiky lines using ODMP for the real galaxy spectrum. Lower: the fitted galaxy spectral function and the estimated variance function using ODMP*

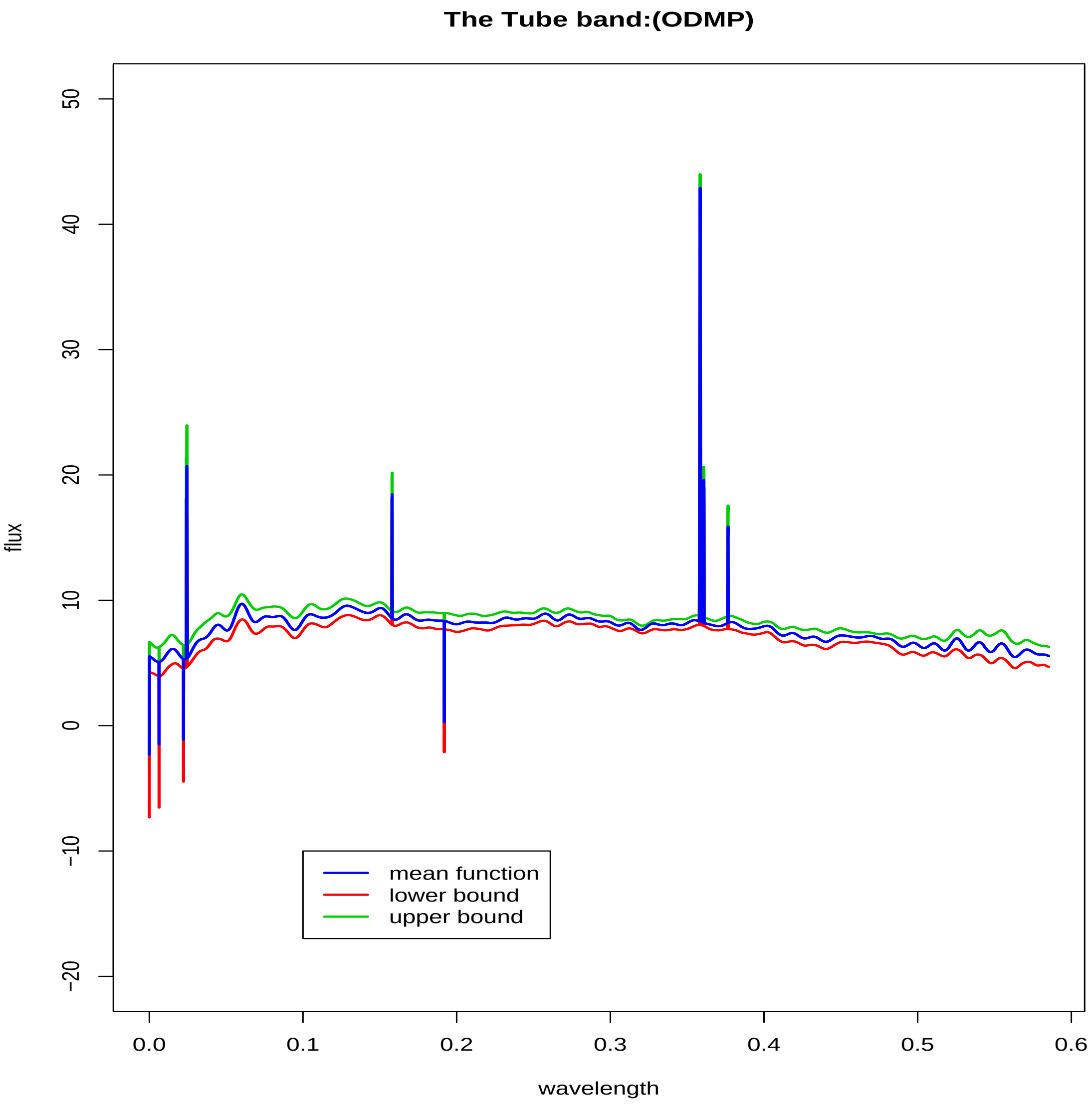


Figure 24: *The simultaneous confidence bands (Tube) and the fitted galaxy spectral function using the ODMP method*

From figure 22 and figure 24, we see that the simultaneous confidence bands are very tight. The Scheffé's bands and the Tube's bands are very consistent. More importantly, from figure 23, we see that using ODMP, our estimated variance function is quite similar to the given variance function. All these suggests a good fit. Also, the fitted continuum and the spiky lines are shown in figure 23.

VII. Conclusions

In this paper, we propose a unified regression framework, named **sparse composite model**, for galaxy spectral function estimation. Under this framework, five different inference methods: ODBP, ODMP, DABP, DAMP, SOMP are developed. Also, different simultaneous confidence bands for degenerated linear models are applied. All these methods are based on basis pursuit or matching pursuit techniques, which have very good theoretical guarantees. More importantly, these methods are very easy to implement and runs efficiently. Using a synthetic galaxy spectrum, all these methods and confidence bands are tested. Based on the simulation, we noticed that two matching pursuit techniques ODMP and SOMP are outperforming the reaming methods. We applied the ODMP on the real galaxy spectrum and obtained a very good fit. Not only the confidence bands are very tight, the estimated variance function is also quite similar to the prior knowledge. More importantly, a big advantage of our approach is that we do not need the prior knowledge or the reference spiky line locations. Since SOMP also work well for the simulated dataset, the next step is to implement the SOMP methods to run on the real datasets.

References

Brutti, P., Genovese, C., Miller, C., Nichol, R. and Wasserman, L. (2005). Spike hunting in galaxy spectra. Tech. Rep. 828, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Carroll, B. and Ostlie, D. (1996). *An Introduction to Modern Astrophysics*. Addison-Weley.

Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20** 33–61.

Claerbout, J. and Muir, F. (1973). Robust modeling of erratic data. *Geophysics* **38** 826–844.

Donoho, D. (1992). Superresolution via sparsity constraints. *SIAM Journal on Mathematical Analysis* **23** 1309–1331.

Donoho, D. (2002). Optimally sparse representation in general (non-orthogonal dictionaries via l1 minimization). *Proceedings of the National Academy of Science (PNAS)* **100** 2197–2202.

Donoho, D. and Elad, M. (2006). On the stability of the basis pursuit in the presence of noise. *Signal Processing* **86** 511–532.

Donoho, D., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory* **52** 6–18.

Donoho, D. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory* **47** 2845–2862.

Donoho, D. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrica* **81** 425–455.

Drew, M. and Funt, B. (1992). Natural metamers. *CVGIP: Image Understanding* **56** 139–151.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *the Annuals of Statistics* **42** 407499.

Elad, M. and Bruckstein, A. (2002). A generalized uncertainty principle and sparse representation in pairs of rn bases. *IEEE Transactions on Information Theory* **48** 2558–2567.

Forsyth, D. A. (1990). A novel algorithm for color constancy. *International Journal of Computer Vision* **5** 5–36.

Forsythe, G. (1957). Generation and use of orthogonal polynomials for data fitting with a digital computer. *Journal of SIAM* **5** 74–88.

Geist, R., Heim, O. and Junkins, S. (1996). Color representation in virtual environments. *COLOR research and application* **21** 121–128.

Gondek, J., Meyer, G. and Newman, J. (1994). Wavelength dependent reflection functions. In *Proceedings of the ACM SIGGRAPH.*

Hall, R. (1989). *Illumination and Color in Computer Generated Imagery.* Springer-Verlag, New York.

Hans, C. and van Dyk, D. (2003). Accounting for absorption line in high energy spectra. *Statistical Challenges in Modern Astronomy III (eds. E. Feigelson and G. Babu)* 429–430.

Heavens, A., Panter1, B., Jimenez, R. and Dunlop, J. (2004). The star-formation history of the universe from the stellar populations of nearby galaxies. *Nature* **428** 625–627.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *the Annuals of Statistics* **32** 15941649.

Lafferty, J. and Wasserman, L. (2005). Rodeo: Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems (NIPS).*

Li, L. and Speed, T. (2000a). Parametric deconvolution of positive spike trains. *the Annuals of Statistics* **28** 1279–1301.

Li, L. and Speed, T. P. (2000b). Deconvolution of sparse positive spikes: is it ill-posed? Tech. Rep. 586, Institute for Pure and Applied Mathematics, UCLA, LA, CA.

Mallat, S. and Zhang, Z. (1994). Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing* **41** 3397–3415.

Marimont, D. and Wandell, B. (1992). Linear models of surface and illuminant spectra. *Journal of the Optical Society of America A: Optics, Image Science, and Vision* **A9** 1905–1913.

Marks, R. (1993). *Advances Topics in Shannon's Sampling and Interpolation Theory.* Springer-Verlag, New York.

Meinshausen, N. (2005). Relaxed lasso. *Computational Statistics and Data Analysis (to appear)* .

Moller, C., v. Alvensleben, U. F. and Fricke., K. (1997). Metallicity indicators across the spectrum of composite stellar populations. *Astronomy and Astrophysics* **317** 686–688.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization.* Springer.

Raso, M. and Fournier, A. (1991). A piecewise polynomial approach to shading using spectral distributions. *Grphics Interface* **91** 40–46.

Saha, P. and Williams, T. B. (1994). Unfolding kinematics from galaxy spectra: A bayesian method. *The Astronomical Journal* **107** 1295–1302.

Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *the Annuals of Statistics* **22** 1328–1345.

Sun, Y. (2000). *A Spectrum-Based Framework for Realistic Image Synthesis.* Ph.D. thesis, Department of Computer Science, Simon Fraser University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288.

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4** 1035–1038.

Tropp, J. (2004). Greedy is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory* **50** 2231–2242.

Tropp, J. (2005). Recovery of short, complex linear combinations via l_1 minimization. *IEEE Transactions on Information Theory* **51** 1568–1570.

Tropp, J. (2006). Just relax: Convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory* **51** 1030–1051.

Tropp, J. and Gilbert, A. C. (2006). Signal recovery from partial information via orthogonal matching pursuit: The gaussian case. Tech. rep., University of Texas-Austin, Austin, TX.

Tropp, J., Gilbert, A. C., Muthukrishnan, S. and Strauss, M. J. (2003). Improved sparse approximation over quasi-incoherent dictionaries. In *Proceedings of the 2003 IEEE International Conference on Image Processing*.

Tropp, J., Gilbert, A. C. and Strauss, M. J. (2006a). Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing* **86** 572–588.

Tropp, J., Gilbert, A. C. and Strauss, M. J. (2006b). Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing* **86** 572–588.

van Dyk, D., Connors, A., Kashyap, V. and Siemiginowska, A. (2001). Analysis of engery spectra with low photon counts via bayesian posterior simulation. *The Astrophysical Journal* **548** 224–243.

van Dyk, D., Connors, A., Kashyap, V. and Siemiginowska, A. (2002). Accounting for absorption lines in images obtained with the chandra x-ray observatory. *Spatial Cluster Modelling (eds. D. Denison and A. Lawson)* 175–198.

Vazdekis, A. and Arimoto, N. (1999). A robust age indicator for old stellar populations. *The Astrophysical Journal* **525** 144–152.

Vrhel, M., Gershon, R. and Iwan, L. (1994). Measurement and analysis of object reflectance spectra. *COLOR research and application* **19** 4–9.

Wakker, B. and van Woerden, H. (1997). High-velocity clouds. *Annual Review of Astronomy and Astrophysics* **35** 217–226.

Wang, H., Li, G. and Jiang, G. (2004). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics (to appear)* .

Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer-Verlag, New York.

Weisberg, S. (2005). *Applied Linear Regression.* Wiley.