



Analytic Marginalization of Absorption Line Continua

Kirill Tchernyshyov^{1,2} 

¹ Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD, USA; ktcherny@gmail.com

² Department of Astronomy, University of Washington, Seattle, WA, USA

Received 2019 January 18; revised 2019 December 5; accepted 2019 December 7; published 2020 January 23

Abstract

Absorption line spectroscopy is a powerful way of measuring properties of stars and the interstellar medium. Absorption spectra are often analyzed manually, an approach that limits reproducibility and which cannot practically be applied to modern data sets consisting of thousands or even millions of spectra. Simultaneous probabilistic modeling of absorption features and continuum shape is a promising approach for automating this analysis. Existing implementations of this approach use numerical methods such as Markov Chain Monte Carlo to marginalize over the continuum parameters. When continua are parameterized as linear functions such as polynomials or splines, it is possible to reduce continuum parameter marginalization to an integral over a multivariate normal distribution, which has a known closed form. Analytic marginalization makes it possible to combine optimization for absorption line parameters with marginalization of nuisance continuum parameters. We compare the accuracy to within which absorption line parameters can be recovered using different continuum placement methods and find that marginalization with an informative prior on continuum parameters is a clear improvement over other continuum placement methods over a broad range of signal-to-noise ratios. We implement analytic marginalization over linear continuum parameters in the open-source package `amlc`.

Unified Astronomy Thesaurus concepts: [Spectroscopy \(1558\)](#); [Astrostatistics techniques \(1886\)](#); [Interstellar line absorption \(843\)](#)

1. Introduction

Absorption lines contain information on the composition and properties of interstellar matter (ISM) and stellar atmospheres. To extract this information, it is necessary to decompose the spectrum into absorption features and the intrinsic flux, typically referred to as the continuum, produced by the illuminating background source toward which the absorption is seen. The most common way of doing this separation has been manually finding regions in a spectrum that do not contain absorption features, fitting a function to these regions, and using this function to interpolate over the absorption features. Given the longevity and popularity of this approach, it is clear that it can produce acceptable results. It does, however, have two important weaknesses. The first is that every spectrum must be examined and interacted with by a human. This cannot efficiently be done for data sets containing thousands or even millions of spectra. The second is that it is unlikely that the absorption parameter estimator this procedure implicitly defines uses data efficiently. There is variance between analyses done by different humans as well as between analyses done by the same human at different times. If there is a subset of analysts whose estimates are the most accurate and precise, then the estimates of the rest are using the available data inefficiently.

An alternative approach is to infer absorption line and continuum parameters simultaneously. To improve the accuracy of the inferred absorption line parameters, it can be useful to marginalize over, rather than fit for, the continuum parameters. This has been done in packages meant for the analysis of absorption lines from both the ISM (`BayesVP`, Liang et al. 2018) and stellar atmospheres (`Starfish`, Czekala et al. 2015; `sick`, Casey 2016). In these packages, continuum parameter marginalization is done numerically, using Markov Chain Monte Carlo (MCMC). As the authors of two of these

packages point out, including large numbers of continuum parameters in MCMC sampling leads to long convergence and autocorrelation times. To keep the number of continuum parameters low, these packages either do not support (`BayesVP`) or advise against (`sick`) including continuum parameters when simultaneously analyzing multiple spectral segments.

In the packages discussed above and in much of the absorption line analysis literature, the continuum is assumed to be a low-order polynomial or spline. Polynomials and splines are nonlinear functions of their x variable, in this case wavelength, but linear functions of their coefficients. For example, a quadratic function, $f(x) = ax^2 + bx + c$, is nonlinear in x but linear in a , b , and c . The same is true of any function that can be expressed as a linear combination of fixed, possibly nonlinear functions of x . This linearity means that it is possible to marginalize over these coefficients analytically if some additional assumptions hold.

Analytic marginalization has several advantages over numerical marginalization. It reduces the dimensionality of the parameter space that would need to be mapped out by an exact³ probabilistic inference method such as MCMC. This reduction is useful because exact probabilistic inference methods tend to operate more efficiently in spaces of lower dimensionality. The continuum parameter-marginalized likelihood and its gradient can also be used as an objective function for optimization. Finally, it makes marginalizing over different continuum parameterizations trivial.

The assumptions required for this particular form of analytic marginalization are: that the continuum can be expressed as a function that is linear in its parameters; that the priors on the parameters of this function are either improper uniform or

³ In the sense of (possibly approximately) exploring the true posterior probability distribution.

multivariate normal; and that residuals from the model are normally distributed. If these assumptions hold, then, given a model for the absorption, the posterior probability distribution function of the continuum parameters is itself a multivariate normal distribution. The first assumption, that the continuum can be expressed as a function that is linear in its parameters, is fulfilled by commonly used functions such as polynomials and splines. One of the options given in the second assumption, that the priors on the parameters of this function are improper uniform, is made implicitly any time a continuum is fit without explicitly defining priors. The third assumption, that residuals from the model are normally distributed, is made whenever a spectrum is analyzed as a set of continuous quantities, such as fluxes, rather than as a set of discrete photon counts.⁴

The key concept is that when a set of absorption line parameters is specified, the continuum parameters can be treated as additive, rather than multiplicative, linear nuisance parameters. Marginalizing over additive linear nuisance parameters simply updates the covariance matrix of the model residuals; see Luger et al. (2017) for an explanation in an astronomical context. This approach to marginalizing over multiplicative linear nuisance parameters has already been used in several astronomical applications, for example for analyzing sparsely sampled radial velocity measurements (Price-Whelan et al. 2017). Models for absorption line spectra have structural features, such as the presence of a line-spread function (LSF), which need to be accounted for to more efficiently compute marginalized likelihoods and likelihood gradients.

In this work, we derive expressions for these quantities that account for these features. This derivation is given in Section 2. We have created a package, `amlc`,⁵ for evaluating these expressions. The package is described in Appendix A.

The performance of continuum marginalization and other continuum placement methods in absorption line analyses is explored with artificial data in Section 3 and, briefly, with actual data in Section 4. We discuss the assumptions, limitations, and prospects of analytic marginalization in Section 5 and conclude in Section 6.

2. Assumptions and Formalism

We assume the following model for a spectrum \mathbf{y} given parameters θ , \mathbf{m} , and \mathbf{b} :

$$\mathbf{y}(\theta) = \mathbf{L} \left(\mathbf{d}(\theta) \odot \left(\boldsymbol{\mu}_m(\theta) + \sum_{i=1}^P \mathbf{a}_{m,i} m_i \right) + \boldsymbol{\mu}_b(\theta) + \sum_{i=1}^Q \mathbf{a}_{b,i} b_i \right) + \boldsymbol{\varepsilon}. \quad (1)$$

The background source emits a continuum, which is expressed as the sum of a mean term, $\boldsymbol{\mu}_m(\theta)$, which may be a nonlinear function of θ and a linear combination of basis elements $\mathbf{a}_{m,i}$ with coefficients m_i . Intervening matter absorbs part of this continuum with the transmittance function $\mathbf{d}(\theta)$. The absorption happens independently at each wavelength. This is indicated by the elementwise product \odot between the transmittance and continuum. Foregrounds, such as sky lines

or instrumental artifacts are, like the continuum, expressed as the sum of a possibly nonlinear mean term, $\boldsymbol{\mu}_b(\theta)$, and a linear combination of basis elements $\mathbf{a}_{b,i}$ with coefficients b_i . The resulting spectrum is convolved with an LSF \mathbf{L} and observed. $\boldsymbol{\varepsilon}$ are the residuals between the observed \mathbf{y} and the LSF-convolved spectrum and are assumed to be normally distributed with mean zero and covariance matrix \mathbf{K} . The length of the observed spectrum is M , the length of the model spectrum before convolution with the LSF is N , the number of continuum basis elements is P , and the number of foreground basis elements is Q .

Collecting the multiplicative (continuum) and additive (foreground) basis elements $\mathbf{a}_{m,i}$ and $\mathbf{a}_{b,i}$ into matrices \mathbf{A}_m and \mathbf{A}_b and converting the transmittance vector $\mathbf{d}(\theta)$ into the diagonal matrix $\mathbf{D}_\theta \equiv \text{diag}(\mathbf{d}(\theta))$,

$$\mathbf{y} = \mathbf{L}(\boldsymbol{\mu}_b(\theta) + \mathbf{A}_b \mathbf{b} + \mathbf{D}_\theta(\boldsymbol{\mu}_m(\theta) + \mathbf{A}_m \mathbf{m})) + \boldsymbol{\varepsilon} \quad (2)$$

$$\equiv \mathbf{L}(\boldsymbol{\mu}_b(\theta) + \mathbf{D}_\theta \boldsymbol{\mu}_m(\theta) + \mathbf{B} \mathbf{c}) + \boldsymbol{\varepsilon}. \quad (3)$$

In the second expression, \mathbf{B} and \mathbf{c} are defined as

$$\mathbf{B} = [\mathbf{D}_\theta \mathbf{A}_m \quad \mathbf{A}_b] \quad \mathbf{c} = \begin{bmatrix} \mathbf{m} \\ \mathbf{b} \end{bmatrix}. \quad (4)$$

We consider two possible priors for the nuisance parameter vector \mathbf{c} , a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Lambda}$ and an improper uniform distribution of

$$p_n(\mathbf{c}) = \mathcal{N}(0, \boldsymbol{\Lambda}) \text{ (normal)} \quad \text{and} \quad p_u(\mathbf{c}) = \prod_{i=1}^{P+Q} Z_i^{-1} \text{ (uniform)}, \quad (5)$$

where Z_i is an arbitrary positive constant.

2.1. Conditional Probability of the Nuisance Parameters

For both priors, the conditional distribution of \mathbf{c} at fixed θ is proportional to a multivariate normal distribution. The mean $\hat{\mathbf{c}}$ of this normal distribution is

$$\hat{\mathbf{c}}_{n/u} = \mathbf{C}_{n/u}^{-1} \mathbf{B}^T \mathbf{L}^T \mathbf{K}^{-1} \mathbf{r}, \quad (6)$$

where \mathbf{r} is the vector of residuals

$$\mathbf{r} = \mathbf{y} - \mathbf{L}(\boldsymbol{\mu}_b(\theta) + \mathbf{D}_\theta \boldsymbol{\mu}_m(\theta)) \quad (7)$$

and $\mathbf{C}_{n/u}$ is

$$\mathbf{C}_n = \boldsymbol{\Lambda}^{-1} + \mathbf{B}^T \mathbf{L}^T \mathbf{K}^{-1} \mathbf{L} \mathbf{B}, \quad (8)$$

if the prior on \mathbf{c} is normal, and

$$\mathbf{C}_u = \mathbf{B}^T \mathbf{L}^T \mathbf{K}^{-1} \mathbf{L} \mathbf{B} \quad (9)$$

if the prior on \mathbf{c} is uniform. The covariance matrix of the conditional distribution of \mathbf{c} is $\mathbf{C}_{n/u}^{-1}$.

The conditional distribution of \mathbf{c} can be used for visualization and predictive checks. The mean of the conditional distribution is also its mode, so $\mathbf{L} \mathbf{B} \hat{\mathbf{c}}$ is the best-fit model for \mathbf{y} at a given value of θ . Samples drawn from the conditional distribution of \mathbf{c} can be used to visualize the effect and extent of nuisance parameter variation.

⁴ Assuming that a spectrum consists of, e.g., fluxes rather than photon counts does not require the assumption that residuals are normally distributed, but normality has been assumed in every such instance known to the author.

⁵ `amlc` is available at <https://github.com/ktchm/amlc>.

2.2. Marginal Likelihood

Assuming the normal prior $p_n(\mathbf{c})$, marginalizing over \mathbf{c} gives

$$p_n(\mathbf{y}|\theta, \mathbf{L}, \mathbf{B}, \mathbf{K}, \Lambda) = \int_{-\infty}^{+\infty} p(\mathbf{y}|\mathbf{c}, \theta, \mathbf{L}, \mathbf{B}, \mathbf{K}, \Lambda) p_n(\mathbf{c}) d\mathbf{c} \quad (10)$$

$$= (2\pi)^{-\frac{M}{2}} \det(\mathbf{K})^{-\frac{1}{2}} \det(\Lambda)^{-\frac{1}{2}} \det(\mathbf{C}_n)^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\mathbf{r}^T \mathbf{K}^{-1}(\mathbf{r} - \hat{\mathbf{r}}_n)\right], \quad (11)$$

where

$$\hat{\mathbf{r}}_{n/u} = \mathbf{L}\mathbf{B}\hat{\mathbf{c}}_{n/u}. \quad (12)$$

If we instead assume the improper prior $p_u(\mathbf{c})$,

$$p_u(\mathbf{y}|\theta, \mathbf{L}, \mathbf{B}, \mathbf{K}) = \int_{-\infty}^{+\infty} p(\mathbf{y}|\mathbf{c}, \theta, \mathbf{L}, \mathbf{B}, \mathbf{K}) p_u(\mathbf{c}) d\mathbf{c} \quad (13)$$

$$= \left(\prod_{i=1}^{P+Q} Z_i^{-1}\right) (2\pi)^{-\frac{M-(P+Q)}{2}} \det(\mathbf{K})^{-\frac{1}{2}} \det(\mathbf{C}_u)^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\mathbf{r}^T \mathbf{K}^{-1}(\mathbf{r} - \hat{\mathbf{r}}_u)\right]. \quad (14)$$

The marginal likelihood p_u will be proper if \mathbf{C}_u is positive definite, which will be the case when $\mathbf{L}\mathbf{B}$ has full rank and $M \geq P + Q$. The marginal likelihood p_n is always proper because \mathbf{C}_n is always positive definite. \mathbf{C}_n is always positive definite because Λ^{-1} is always positive definite and $\mathbf{B}^T \mathbf{L}^T \mathbf{K}^{-1} \mathbf{L}\mathbf{B}$ is always at least positive semidefinite.

2.3. Gradients

We give expressions for the gradients of $\log(p_n)$ and $\log(p_u)$ with respect to $\mathbf{d}(\theta)$, $\boldsymbol{\mu}_b(\theta)$, and $\boldsymbol{\mu}_m(\theta)$. The gradient of $\log(p)$ with respect to the parameters θ can be obtained by evaluating each of these gradients, computing the Jacobians of $\mathbf{d}(\theta)$, $\boldsymbol{\mu}_b(\theta)$, and $\boldsymbol{\mu}_m(\theta)$ with respect to θ , and applying the chain rule.

The gradient of $\log(p)$ with respect to $\mathbf{d}(\theta)$ is

$$\begin{aligned} \nabla \log(p)(\mathbf{d}(\theta)) &= (\mathbf{L}^T \mathbf{K}^{-1}(\mathbf{r} - \hat{\mathbf{r}}_{n/u})) \odot (\mathbf{B}'\hat{\mathbf{c}} + \boldsymbol{\mu}_m) \\ &\quad - \frac{1}{2}((\mathbf{C}_{n/u}^{-1} \mathbf{B}'^T) \odot (\mathbf{B}^T \mathbf{L}^T \mathbf{K}^{-1} \mathbf{L})) \\ &\quad + (\mathbf{C}_{n/u}^{-1} \mathbf{B}^T \mathbf{L}^T \mathbf{K}^{-1} \mathbf{L}) \odot \mathbf{B}'^T \mathbf{1}, \end{aligned} \quad (15)$$

where $\mathbf{1}$ is a column vector of ones of length $P + Q$. \mathbf{B}' is the sum of derivatives of \mathbf{B} with respect to each element of $\mathbf{d}(\theta)$:

$$\mathbf{B}' = \sum_{i=1}^N \frac{\partial \mathbf{B}}{\partial d_i(\theta)} \quad (16)$$

$$= \sum_{i=1}^N [\mathbf{J}^{i,i} \mathbf{A}_m \quad \mathbf{0} \times \mathbf{A}_b] \quad (17)$$

$$= [\mathbf{A}_m \quad \mathbf{0}], \quad (18)$$

where $\mathbf{J}^{i,i}$ is a square matrix whose (i, i) -th entry is 1 and whose other entries are all 0. The first row of Equation (15) is the gradient of the argument of the exponentials in Equations (10) and (13). The second row is the gradient of $\log(\det(\mathbf{C}_{n/u}))$.

The gradient of $\log(p)$ with respect to $\boldsymbol{\mu}_m(\theta)$ is

$$\nabla \log(p)(\boldsymbol{\mu}_m(\theta)) = \mathbf{D}_\theta \mathbf{L}^T \mathbf{K}^{-1}(\mathbf{r} - \hat{\mathbf{r}}_{n/u}) \quad (19)$$

and the gradient of $\log(p)$ with respect to $\boldsymbol{\mu}_b(\theta)$ is

$$\nabla \log(p)(\boldsymbol{\mu}_b(\theta)) = \mathbf{L}^T \mathbf{K}^{-1}(\mathbf{r} - \hat{\mathbf{r}}_{n/u}). \quad (20)$$

2.4. Marginalizing over Parameterizations

If there are multiple possible models whose marginal likelihoods are available in closed form, it is trivial to marginalize over the choice of model. In the case of continuum marginalization, the models could be polynomials of different degrees. For marginalization over model choice to be well defined and meaningful, the prior within each of the possible models must be proper. It is also necessary to specify a prior over the choice of model. If the models are assumed to be equally likely before seeing the data, the prior probability of each model would be the inverse of the number of possible models. In general, the prior will be a set of weights, one per model, that sum to one.

Given a parameter set θ and T possible continuum bases and priors, the parameter and parameterization-marginalized likelihood is the weighted sum of the parameter-marginalized likelihoods

$$p(\mathbf{y}|\theta, \mathbf{L}, \mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_T, \mathbf{p}, \mathbf{K}, \Lambda_1, \Lambda_2, \dots, \Lambda_T, \mathbf{p}) = \sum_{i=1}^T p_i \times p(\mathbf{y}|\theta, \mathbf{L}, \mathbf{B}_i, \mathbf{K}, \Lambda_i), \quad (21)$$

where the weights p_i are the prior probabilities of each model. This way of marginalizing over models is not specific to the likelihood function described in this section—it works for any set of proper marginalized likelihoods. It cannot be applied if the prior over any of the marginalized parameters is improper (e.g., the uniform prior defined in Equation (5)).

3. Tests on Artificial Data

In this section, we explore how different continuum placement methods affect the accuracy and precision with which column densities can be measured from absorption lines. We do this by generating artificial spectra containing absorption lines with known input parameters and attempting to recover these parameters using different continuum placement methods. The central question of this section is: are line parameters obtained by marginalizing over continuum parameters more accurate and precise than line parameters obtained using other continuum placement strategies? The answer depends on whether or not a correct, informative prior over continuum parameters is used for marginalization. If an informative and accurate prior is available, continuum marginalization produces results that are almost as precise and accurate as those obtained when each test spectrum's true continuum is known. If continuum marginalization is instead done with a diffuse, uninformative prior, the results are no better than simultaneously fitting for line and continuum parameters.

To isolate the effect of continuum placement, we keep the test problem simple: a single resolved and unsaturated absorption line superimposed on a continuum that is a first or second degree polynomial (i.e., a line or a quadratic function). We vary the depth of the absorption line, the extent of the spectrum surrounding the line, and the signal-to-noise ratio (S/N) of the artificial data. We generate 2000 spectra for each combination of total optical depth, S/N, spectrum extent, and

Table 1
Parameters That Define an Artificial Test Spectrum

Parameter	Values
Continuum degree	1, 2
Total optical depth	1.0, 2.0, 3.0
Velocity extent	35, 50, 65 km s ⁻¹
Signal-to-noise ratio	5, 6.66..., ..., 20; $\Delta S/N = 5/6 = 1.66\dots$
Continuum coefficients	Randomly generated
Noise	Randomly generated

Note. Multiple realizations are generated for fixed values of the degree of the polynomial describing the continuum, total optical depth of the absorption feature, extent of the spectrum, and signal-to-noise ratio of the spectrum. Continuum coefficients and the noise vector are randomly generated for each realization.

continuum degree. Each spectrum has a different set of continuum parameters, which are generated from a normal distribution. The input parameters we adopt are listed in Table 1. Example spectra generated using each of the considered total optical depths and extents are shown in Figure 1.

3.1. Continuum Placement Methods

We consider two categories of continuum placement method: ones in which continuum parameters are optimized for, or fitted continuum (FC) methods, and ones in which continuum parameters are marginalized over, or marginalized continuum (MC) methods. We also consider a case where the true continuum is known (case TC) and only the line parameters need to be determined. Since case TC’s error in line parameters due to continuum placement error is, by definition, identically zero, case TC provides an estimate of the error in line parameters purely due to independent normally distributed noise. To determine line parameters from a spectrum in case TC, we minimize⁶ the discrepancy between the data and a model consisting of the correct continuum attenuated by an absorption line.

Within the two categories, we define methods based on different amounts of prior knowledge about the continuum. In the continuum fitting category, we provide two levels of prior knowledge: knowledge of which part of the spectrum is effectively free of absorption and no additional knowledge besides which continuum parameterization to use (i.e., whether to use a first or second degree polynomial). The first level represents a perfectly competent analyst or algorithm selecting an absorption-free region over which to fit a continuum. The second level provides enough information to ensure that the correct solution is allowed by the model. To determine line parameters given a true absorption-free region, we fit a continuum to this region and then optimize for line parameters assuming the obtained continuum. To determine line parameters just given the correct continuum parameterization, we simultaneously optimize for continuum and line parameters. We will refer to these continuum placement methods by the abbreviations TR–FC (true region–fit continuum) and FC (fit continuum).

We provide three levels of prior knowledge in the continuum marginalization category: knowledge of the correct prior; knowledge of several possible priors, one of which is the correct prior;

and no meaningful prior knowledge. The correct prior is the distribution used to generate the continuum parameters. This level of knowledge provides a baseline for how well it is possible to recover absorption line parameters while using continuum marginalization. The second level can be thought of as having accurate knowledge of the range of plausible continuum shapes in several distinct classes of continuum, but not knowing in advance which class should be used for a given spectrum. The third level is a pessimistic scenario and is analogous to method FC. To ensure numerical stability, we still use a proper prior in this method, but one which is orders of magnitude broader than the correct one. In the first and third methods, absorption line fitting is done using the logarithm of Equation (10) as the objective function. In the second method, the objective function is the logarithm of Equation (21). We refer to these continuum placement methods by the abbreviations TCov–MC (true covariance–marginalized continuum), MCov–MC (marginalized covariance–marginalized continuum), and DCov–MC (diffuse covariance–marginalized continuum).

3.2. Results

Using these six objective functions and nonlinear optimization routines from the SciPy package, we fit absorption profiles to each of the generated test spectra. This yields one set of line parameters—center, breadth, and total optical depth—per test spectrum. Because it is the most challenging line parameter to correctly recover, we will focus on the total optical depth. To quantify the quality of recovery, we compute the root mean square error (RMSE) of the logarithm of the total optical depth:

$$\text{RMSE}(\log \tau) = \sqrt{\frac{1}{N} \sum_{i=1}^N (\log \hat{\tau}_i - \log \tau_{\text{true}})^2}. \quad (22)$$

$\hat{\tau}_i$ is the total optical depth value recovered from spectrum realization i and τ_{input} is the total optical depth actually used to generate the spectra. We use the RMSE of the logarithm instead of the linear value of the total optical depth because it is equal to the RMSE of a corresponding column density. The total optical depth of a line is equal to the product of the column density of the absorber and some physical constants. The difference of the logarithms of a pair of total optical depths is therefore equal to the difference of the logarithms of the corresponding column densities—the multiplicative constant factors become additive and cancel.

The RMSEs for each combination of input parameter set and continuum placement method are shown in Figures 2 and 3. One way to summarize these results across the tested parameter space is to compare the performance of each method relative to the performance of each other method. To do this quantitatively, we calculate the fraction of cases in which the RMSE of method A is less than or equal to the RMSE of method B. A visual representation of this relative performance summary is shown in Figure 4.

The two informed continuum marginalization methods, TCov–MC and MCov–MC, consistently yield the lowest RMSEs of all the continuum placement methods. For about 80% of the fixed input parameter combinations, TCov–MC and MCov–MC perform as well as the known-continuum reference case, TC. Informed continuum fitting, TR–FC, performs more poorly than informed marginalization but better than both

⁶ All optimization is done using routines from the SciPy package’s optimize module.

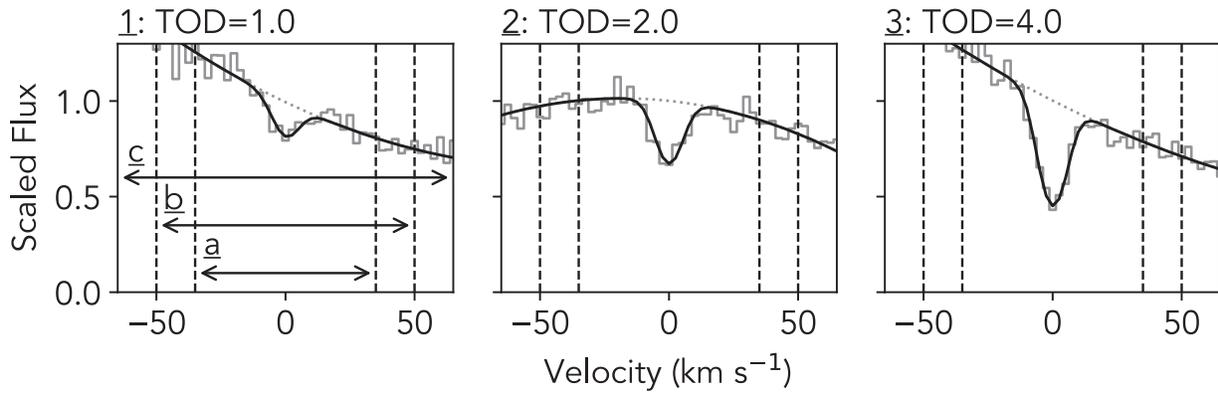


Figure 1. Examples of the artificial spectra that are used in the tests described in Section 3. Each panel shows a noise-free spectrum with absorption (solid black line), a noise-free continuum (dotted gray line), and a noisy realization of the spectrum with absorption (solid gray). These artificial spectra have a signal-to-noise ratio of 20 and a polynomial of degree 2 as the continuum. The panels are numbered in order of increasing total optical depth (TOD). The vertical lines and letters shown in each panel correspond to different spectrum extents, ± 35 (a), ± 50 (b), and ± 65 km s^{-1} (c). The numbers and letters used to indicate the TODs and spectrum extents in this figure are consistent with the labeling used in Figures 2 and 3.

uninformed fitting and uninformed marginalization. Finally, uninformed fitting, FC, performs better on average than uninformed marginalization, DCov-MC. Below, we discuss a few aspects of the results in more detail.

3.2.1. The Value of Different Kinds of Prior Continuum Information

All of these methods combine prior and observed continuum information. The performance of the methods reflects the value of the prior information available to them and the efficiency with which they extract continuum information from the observed spectrum.

By definition, case TC has as much continuum information as it is possible to have. The fact that methods TCov-MC and MCov-MC perform almost as well as case TC suggests that the accurate priors used in these methods are almost as informative as knowing the actual continuum. These particular priors are not so constraining that all continuum realizations would be indistinguishable. The (clearly distinguishable) continua of the spectra shown in Figure 1 were generated from the prior used in method TCov-MC. The prior used by method MCov-MC is less specific, but apparently no worse, than that of TCov-MC. The fact that these priors are so informative is noteworthy because it is, in many cases, possible to infer them; this point is discussed further in Section 5.2.

The performance of method TR-FC suggests that knowing where the spectrum is free of absorption has value, but not as much as an accurate prior on the continuum shape. The value of this information is greatest for parameter sets with the lowest TOD, i.e., the least contrast between the continuum and absorption.

3.2.2. Solutions with RMSEs that are Lower than those of the TC Solution are Biased

There is a small number of fixed input parameter combinations for which the RMSE of method FC is lower than that of case TC. All of these combinations have low S/Ns, small TODs, and short continuum extents. This can be explained by the fact that much of the RMSE of method FC for these combinations is due to bias, rather than variance. Method FC is consistently finding solutions that are incorrect by an amount that is small relative to the scatter in the less-biased TC solution. This bias is favorable for those particular true line parameters, but may be unfavorable for other values of the true line parameters.

3.2.3. Continuum Marginalization is not Substantially Slower than Continuum Fitting

On a single core of a 2.2 GHz Intel i7 processor, obtaining line parameters for 2000 spectra took approximately 2 s in case TC, 5 s with method TR-FC, 10 s with methods MCov-MC and DCov-MC, 15 s with method FC, and 20 s with method MCov-MC. Method TR-FC is a factor of 2–4 faster than the other continuum placement methods once absorption is masked. When analyzing actual observations, this advantage would be outweighed by the human interaction time required to define absorption masks. More generally, all of these times are short enough that any spectrum-by-spectrum interaction by a human would be the main processing bottleneck.

4. Demonstration on Actual Data

In this section, we analyze absorption features using continuum marginalization in a case where all continuum placement methods should agree. The absorption features are due to C I in the Large Magellanic Cloud (LMC). The data are spectra of the LMC stars Sk-67 5, Sk-68 73, and Sk-70 115 taken with the Space Telescope Imaging Spectrograph (STIS; Woodgate et al. 1998) on board the *Hubble Space Telescope* (*HST*) using the E140H grating ($R \sim 114,000$). These absorption features in this data set were previously analyzed by Welty et al. (2016), whose C I results we adopt as ground truth. Welty et al. (2016) did continuum placement by fitting polynomials to manually selected line-free regions.

We downloaded the default pipeline-extracted one-dimensional spectra from the Mikulski Archive for Space Telescopes. Individual exposures within a single echelle order were shifted to the wavelength grid of one of the exposures using nearest-neighbor interpolation and coadded. We did not splice different echelle orders into a combined spectrum.

To measure C I column densities, we fit Voigt profiles to the C I, C I*, and C I** absorption in these spectra. We modeled the continuum as the sum of a first degree polynomial and a Gaussian process with a Matern-5/2 kernel. The LSFs we used were downloaded from the *HST*-STIS website.⁷ Following the example of Welty et al. (2016), we analyzed only the C I multiplets near 1280.1 Å and 1328.8 Å. All absorption

⁷ <http://www.stsci.edu/hst/instrumentation/stis/performance/spectral-resolution>

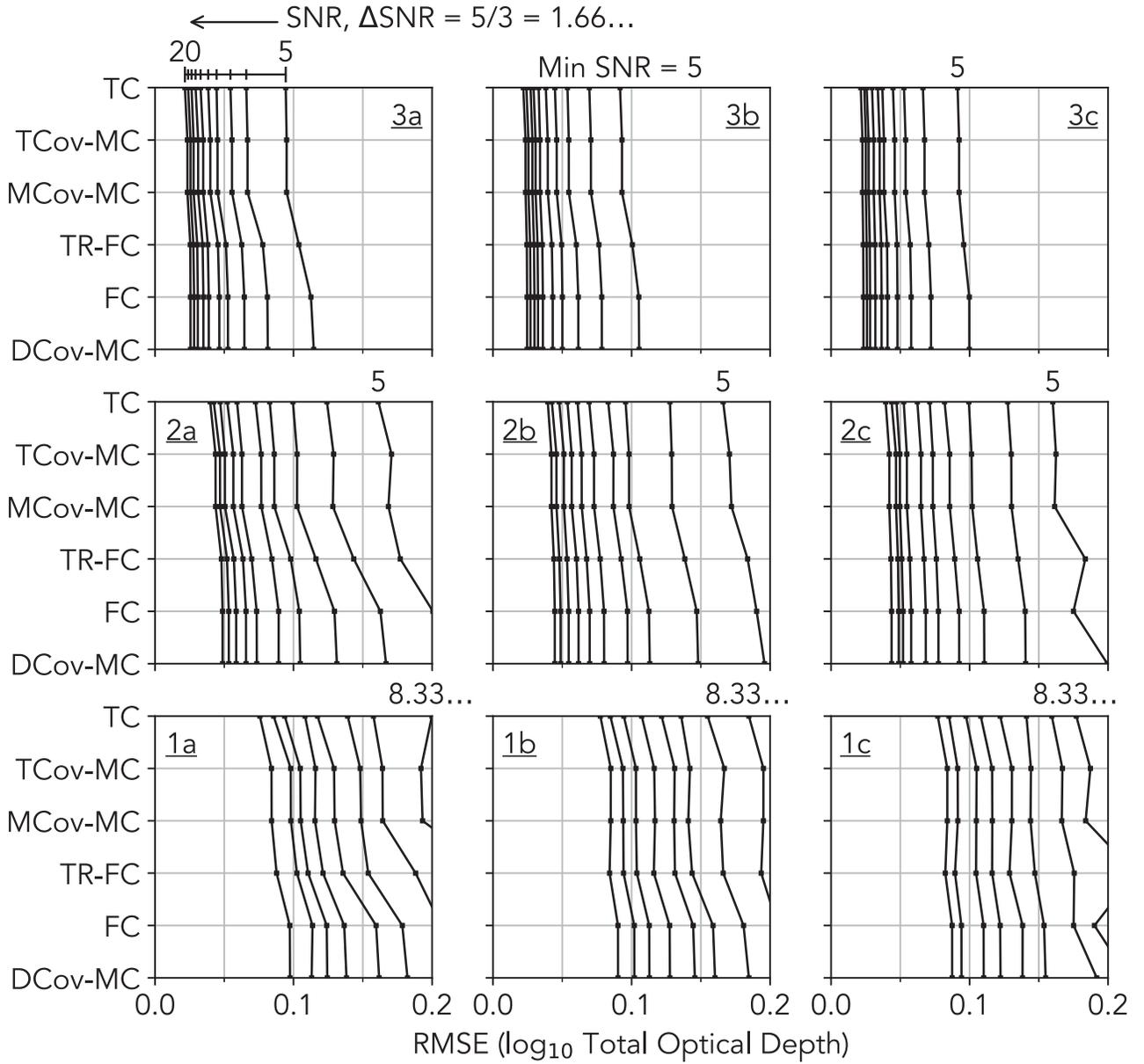


Figure 2. How the root mean square error (RMSE) in the recovered log total optical depth (TOD) varies as a result of changing the true TOD of the absorption feature, the extent of the spectrum around the absorption feature, the signal-to-noise ratio (S/N) of the spectrum, and the continuum placement method. In this figure, the continuum is a polynomial of degree 1, i.e., a line. Each panel corresponds to a different combination of true TOD and spectrum extent. The TOD increases from bottom to top and is indexed by the underlined number in the corner of each panel; the indices 1, 2, and 3 correspond to the linear (i.e., non-log) TODs of 1.0, 2.0, and 4.0. The spectrum extent increases from left to right and is indexed by the underlined letter in the corner of each panel; the indices a, b, and c correspond to extents of ± 35 , ± 50 , and ± 65 km s^{-1} . These labels are consistent with those used in Figure 1, which shows examples of spectra at each combination of TOD and spectral extent. Within each panel, each line connecting a series of points corresponds to a different S/N; these are labeled at the top of panel (3a). The minimum S/N whose RMSEs are small enough to fall within the range shown is indicated at the top of each panel. The vertical axis within each panel corresponds to different continuum placement methods. The full names corresponding to the abbreviations shown here are given in Section 3.1.

features for a target were fit simultaneously by maximizing the continuum-marginalized likelihood function of the data given the absorption model. Each velocity component had a single central velocity and breadth and a separate column density for the ground and two excited states of C I. Optimizing over the profile parameters while marginalizing over continuum parameters for a single object took between a few and 10 s per target. The computation time is dominated by evaluating the continuum-marginalized likelihood and its gradient with respect to the Voigt profile parameters.

The maximum likelihood absorption solution for target Sk-67 5 is shown in Figure 5. For all three targets, the total

recovered C I column densities across all velocity components and excitation levels agree within the uncertainties of those found in Welty et al. (2016). This agreement is expected, since these are data with high spectral resolution and S/N and the absorption lines are narrow and clearly distinct from the continuum. In this regime, all valid continuum placement methods should perform equally well.

We have chosen a data set with these characteristics to limit the plausible reasons for any potential disagreement to the continuum placement method. Preparing a more challenging low S/N spectrum for analysis often requires bespoke, and proprietary, processing (see, e.g., Wakker et al. 2015 for a

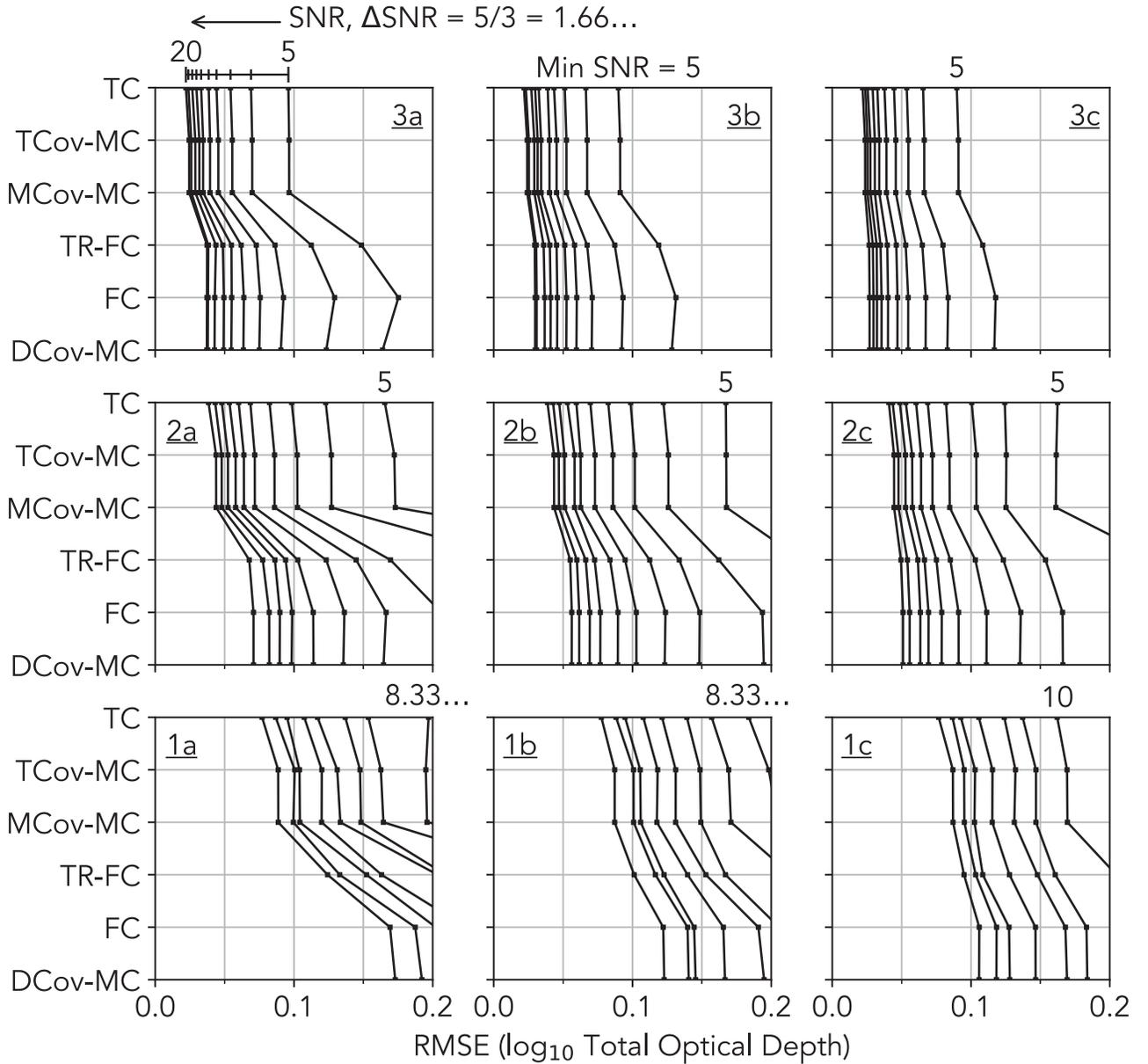


Figure 3. How the root mean square error (RMSE) in the recovered log total optical depth (TOD) varies as a result of changing the true TOD of the absorption feature, the extent of the spectrum around the absorption feature, the signal-to-noise ratio (S/N) of the spectrum, and the continuum placement method. In this figure, the continuum is a polynomial of degree 2, i.e., is a quadratic function. Panel labels and axes are defined in the caption of Figure 2.

well-documented example). As a result, there is no guarantee that there should be agreement between line parameters obtained using the same continuum placement and line fitting method applied to different reductions. To be able to compare results produced by continuum marginalization against the best efforts of other analysts, we therefore chose a data set where there should be no such ambiguity.

5. Discussion

5.1. Assumptions and Consequences

The explicit assumptions of the analytic marginalization method are that the continuum is a linear combination of basis functions, that the prior on the coefficients of this linear combination is the improper uniform or multivariate normal distribution, that residuals between the data and model are normally distributed, and that the covariance matrix of the

residuals does not depend on the continuum. It is obvious that these assumptions do not hold in a strict sense for any data set. For example, both of these priors do not place any hard constraints on the continuum coefficients, meaning that negative continuum values are allowed by the model. This is unphysical, but practically not an issue in most applications—a negative continuum would be inconsistent with a typical spectrum and would therefore have a low likelihood. In addition to these largely irrelevant violations of the strict letter of the assumptions, there are also some meaningful cases.

One such case is data in the low photon count regime. Depending on whether a spectrum is left in terms of photons or converted to a rate (or further transformed into, e.g., a flux) and on whether and how background subtraction is done, the value of a low photon count measurement is best described by a Poisson, Skellam, Gamma, or difference-of-Gammas distribution. All of these distributions converge to the normal

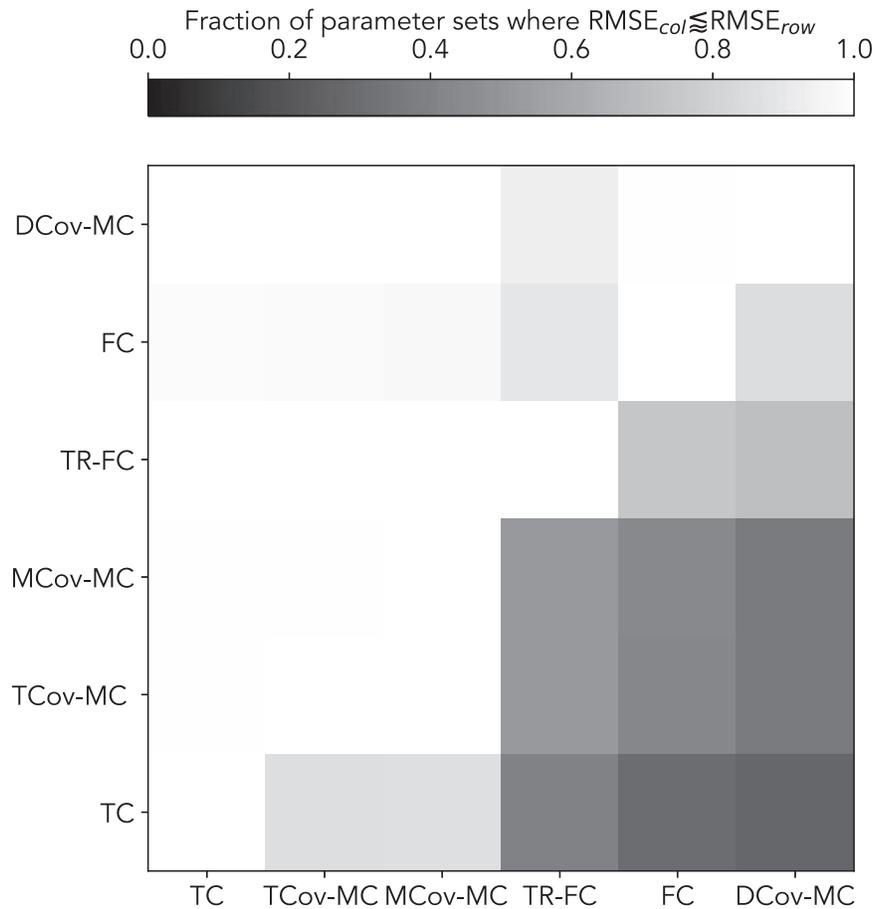


Figure 4. Summarized comparison of the relative performance of different continuum placement methods. Each row and each column correspond to a different method: knowing the true continuum (TC); marginalization with the correct prior over continuum parameters (TCov-MC); marginalization with an prior over continuum parameter priors that include the correct prior (MCov-MC); continuum fitting to a correctly chosen absorption-free portion of the spectrum (TR-FC); simultaneously fitting continuum and absorption parameters (FC); and marginalization with an uninformative prior (DCov-MC). These methods are described in greater detail in Section 3.1. The color of each pixel indicates the fraction of input parameter sets for which the log total optical depth (TOD) root mean square error (RMSE) of the method on the x -axis is less than or within 0.01 dex of the log TOD RMSE of the method on the y -axis. This comparison quantity is an average across the different continuum degrees, true TODs, and spectrum extents and signal-to-noise ratios. The higher the fraction is, the better—a value of 1.0 means that the method on the x -axis is at least as precise/accurate as the method on the y -axis for every input parameter set. Methods TCov-MC and MCov-MC perform as well as knowing the true continuum (case TC) in approximately 80% of our test setups and outperform all three of the other methods in almost all of our test setups. Method DCov-MC has the worst performance of all of the considered methods.

distribution as the number of photons grows, but are poorly approximated by it at low counts. Furthermore, the variance of these distributions is a function of the true count rate. Point estimates of the uncertainties derived from the observed number of counts or a related quantity will therefore themselves be uncertain. Errors in the measurement itself and the estimate of its uncertainty will also be covariant. Assuming normality and using uncertainty point estimates in this regime can produce biased estimates of the parameters. This means that analytic marginalization of the kind described in this work should not be applied to low S/N X-ray or UV spectra.

A more implicit nontrivial assumption that can be broken is that the absorption model is realistic. Absorption features that cannot be described by the absorption model will be described by the continuum model. For example, if a region of a spectrum contains two clearly distinct absorption lines but the model only allows for a single line, the presence of the unmodeled line will bias the continuum model. In short, improvements in continuum modeling cannot solve problems of absorption model misspecification. However, the model does not need to be realistic in an absolute physical sense. If the two hypothetical lines appear to be and can be precisely emulated

by a single line at the resolution of the spectrum, then an absorption model consisting of a single line will suffice for the purpose of avoiding bias in the continuum model.

Another nontrivial assumption is that the continuum can be described by an effective, rather than a physical, model. The continua of most background sources that are used for absorption spectroscopy can be approximated in this way. Examples of sources with slowly varying continua include quasars and (particularly rapidly rotating) hot stars. With flexible linear models such as Gaussian processes, it is even possible to describe more complicated pseudo-continua. To describe sources such as cool stars, however, it is still necessary to use a nonlinear model. Marginalizable linear models can still be useful even in this case as a way of introducing small corrections for pseudo-continuum features that are not perfectly described by the nonlinear model.

5.2. Informative Priors Are Valuable and Realistically Obtainable

The continuum placement methods with the best absorption line parameter measurement performance in Section 3 were

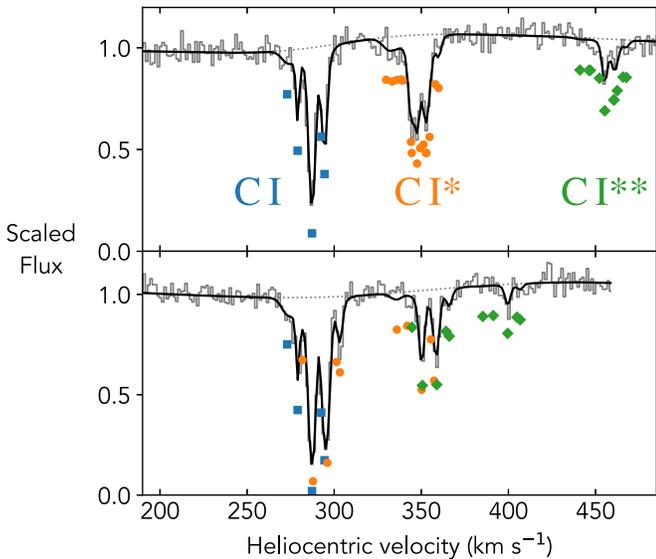


Figure 5. A fit to absorption from ground and excited state C I absorption toward the star Sk-67 5 in the Large Magellanic Cloud. The data, continuum, and combined fit are shown in solid gray, dotted gray, and solid black lines. The velocities shown on the x-axis are relative to the C I $\lambda 1329$ Å (top panel) and $\lambda 1277$ Å lines (bottom panel). Line centers of the ground and first two excited states are indicated by data points of different shapes and colors. The number of excited state line centers is greater than the number of ground state line centers because there are multiple excited state transitions and a single ground state transition in each velocity range.

both forms of continuum marginalization with an accurate and informative prior on continuum parameters. In one case, the prior was the distribution that had actually been used to generate the test continua. In the other, the prior was a mixture of several different priors, one of which was the correct one. Both methods performed equally well and, for 80% of input parameter combinations, matched the accuracy and precision obtained by fitting for line parameters using the true, input continuum.

While it is not possible to know the true continuum of a real background source, it is, in many cases, possible to learn an accurate and informative prior on continuum parameters. The main requirement is the availability of a training set: multiple continuum observations with consistent statistical properties. This requirement can be fulfilled using background sources of a single class whose member-to-member continuum variation can be described using a small number of parameters. Quasars are one such class. A number of authors have derived bases for describing quasar continua (e.g., Suzuki 2006; Zhu & Ménard 2013). These bases provide informative priors on continua even when the priors on their coefficients are uninformative. Eilers et al. (2017) adopt one of these bases and numerically marginalize over continuum coefficients as part of an analysis of absorption in the Lyman α forest.

The spectrum of a single object can also serve as a training set for a prior that is applied over small segments of that object’s spectrum. If the goal is to analyze absorption over regions spanning 200 km s^{-1} , for example, the procedure would be to select other 200 km s^{-1} regions with statistically similar continua, use these regions to derive a basis and prior, and then apply that prior in the analysis of the region of interest. This procedure depends on there being statistically consistent regions of continuum in the spectrum of the object in question.

5.3. Applications of Analytic Marginalization

Section 3 shows that marginalization over continuum parameters and parameterizations allows measurements of absorption line parameters that are as precise and accurate as measurements obtained with knowledge of the true continuum so long as accurate and informative prior information is available. Without an informative prior, analytical marginalization is just a potentially more computationally efficient replacement for numerical marginalization in probabilistic inference schemes; a quantitative assessment of possible efficiency gains in MCMC is presented in Appendix B. The two scenarios discussed above in which it is possible to infer an informative prior may be useful in two categories of the absorption line analysis problem.

The first scenario, in which the background source belongs to a class with limited variation, is appropriate for modern spectroscopic surveys that produce thousands to millions of spectra. Surveys such as the Galactic Archaeology with HERMES (GALAH; Buder et al. 2018) observe sources whose continua do not all have consistent statistical properties. However, it is possible to divide these sources into multiple categories such that there is limited variation within each category. In the case of GALAH, such a division could be based on stellar parameters. The spectra in each division would have their own prior over continuum shapes. Absorption features in these spectra would need to be analyzed using optimization methods rather than probabilistic ones due to computation time considerations.

The second scenario, in which the prior is instead learned from absorption-free portions throughout the spectrum, is appropriate for detailed analyses of individual objects. Splitting the spectrum up to produce a training set would require either human interaction or sophisticated and time-consuming inference, especially in cases where the statistical properties of the continuum vary across the spectrum. The UV spectrum of a rapidly rotating O star, for example, can contain regions that are largely devoid of features, regions containing many broad stellar absorption lines, and regions dominated by stellar winds. If the absorption features of interest are present in all of these regions, it would be necessary to infer a continuum shape prior for each region. Absorption features in these spectra could be analyzed using probabilistic methods such as MCMC.

5.4. Toward Automation

Continuum marginalization can be used in a way that brings the analysis of absorption spectra closer to automation. In the analysis procedure described in the discussion of large surveys immediately above the present section, for example, the role of a human analyst in continuum placement would be limited to the division of spectra into categories based on stellar parameters. However, this procedure does not account for another step of absorption line analyses that typically requires human intervention—specifying the absorption model. This includes deciding on a number of components into which an absorption feature should be split.

For large surveys, a derivative spectroscopy technique along the lines of autonomous Gaussian decomposition (AGD) could be a viable solution (Lindner et al. 2015; Riener et al. 2019). If the absorption lines in question are not saturated and can be described as Gaussians, AGD itself may be sufficient. For individual spectra, where MCMC is possible, component

structure specification can be done using trans-dimensional inference, in which the dimensionality of parameter space (in this case the number of sets of absorption line parameters) is itself a parameter of the model.

6. Conclusion

Absorption lines are an important source of information about stars and the gaseous universe. As larger spectroscopic data sets become available and as reproducibility becomes more standard in astronomy, it becomes necessary to move beyond ad hoc absorption line analysis methods, particularly ones in which a human directly interacts with data. In multiple recent works, there have been attempts to partially automate continuum placement by including and marginalizing over continuum parameters in probabilistic spectral models. Marginalizing over continuum parameters has, in these works, been hypothesized, though not experimentally shown, to also improve the accuracy of the recovered absorption line parameters. Despite these advantages, this approach has so far not become popular, in part due to the computational expense of numerically marginalizing over these additional parameters.

In this work, we have shown that it is possible in many cases to replace this numerical marginalization with analytic marginalization (Section 2). Marginalizing over different possible continuum models, as well as over the parameters of each individual possible model, is a trivial extension of this result. Using tests on artificial data, we have shown that when an accurate and informative prior on continuum parameters is available, marginalizing over continuum parameters in an absorption line analysis produces absorption line parameters that are as accurate and precise as line parameters obtained when the true continuum is known. This is true even at low S/N where other continuum placement methods do not perform as well.

We have released an open-source python package, `amlc`, which can be used to evaluate continuum parameter-marginalized likelihoods and related quantities. Features of this package are described in Appendix A. It is meant to be used as a drop-in replacement for likelihood functions in existing absorption spectrum analysis tools.

The author thanks the referee for comments and suggestions that greatly improved this work, Andrew Casey, Andrew Fox, Cameron Liang, and Yong Zheng for discussions about use cases, and Joshua Peek and Linda Tchernyshyov for helpful comments. This research is based on observations made with the NASA/ESA *Hubble Space Telescope* obtained from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 526555. These observations are associated with programs 9757 and 12978. K.T. was supported by the National Science Foundation under grant 1616177.

Software: `emcee` (Foreman-Mackey et al. 2013), `matplotlib` (Hunter 2007), `numpy` (van der Walt et al. 2011), `scipy` (Jones et al. 2001).

Appendix A Implementation and Demonstration

In this Appendix, we describe how `amlc` is implemented (Appendix A.1), list some of its capabilities (Appendix A.2),

and show how the computation time of different calculations grows with data set and continuum model size (Appendix A.3).

A.1. Implementation

We have implemented `amlc` as a pure Python package with `numpy` and `scipy` as dependencies. `amlc` does not contain functionality for building LSFs or computing transmittances from absorption parameters and is not intended to be a stand-alone analysis tool. It is meant to be used as a drop-in likelihood function replacement in analysis packages or scripts.

A.2. Package Functionality

This package was designed for a use case where the log-marginal likelihood and its gradient are evaluated at many different values of the θ -dependent parameters (see Section 2) while the θ -independent parameters are held constant. The core feature of the package is the `MarginalizedLikelihood` class. A `MarginalizedLikelihood` instance stores θ -independent parts of the model and pre-computes quantities that are reused during repeated marginalized likelihood evaluations. In particular, it stores the data covariance matrix \mathbf{K} ; the c prior covariance matrix $\mathbf{\Lambda}$ and its explicit inverse, if applicable; and the LSF mapping \mathbf{L} and its transpose.

Both covariance matrices can be diagonal or fully general. The package includes the `CovarianceMatrix` class, which defines a consistent interface for calculations, and two subclasses, `DiagonalCovarianceMatrix` and `GeneralCovarianceMatrix`. `DiagonalCovarianceMatrix` wraps the simple, one-dimensional determinant and inverse calculations possible with a covariance matrix consisting purely of variances and does the book-keeping required to produce output with the correct shape. `GeneralCovarianceMatrix` uses the Cholesky decomposition of the supplied covariance matrix to calculate its determinant and to left multiply matrices and vectors by its inverse. Computing the Cholesky decomposition of a general covariance matrix of size M by M takes $\mathcal{O}(M^3)$ calculations, making it prohibitively computationally expensive for large M .

The LSF mapping \mathbf{L} can be any object that implements the matrix multiplication interface, i.e., has a `matmul` or `__matmul__` method. For example, \mathbf{L} can be a dense matrix represented by a `numpy` array, a sparse matrix represented by a `scipy.sparse` matrix, or a convolution operator represented by a `scipy.sparse.linalg.LinearOperator`. \mathbf{L} can also be the identity mapping (indicated by `None`), in which case it is left out of any likelihood calculations.

A.3. Computation Time as a Function of Data Set and Basis Size

The most time-consuming step in computing all of the quantities derived in Section 2 is forming the matrix $\mathbf{C}_{n/u}$. This step requires matrix–matrix products, while most other steps only involve matrix–vector products. These expensive products are $\mathbf{L}\mathbf{B}$ and $\mathbf{K}^{-1}(\mathbf{L}\mathbf{B})$. The amount of time required to compute these products depends on the structures \mathbf{L} and \mathbf{K} .

\mathbf{L} can be the identity matrix, a dense matrix, a sparse matrix, or a linear mapping such as convolution. The fastest case is when \mathbf{L} is the identity matrix, since then $\mathbf{L}\mathbf{B}$ does not need to be computed. The slowest case is when it is a dense matrix, in which case computation time grows as $\mathcal{O}(MN(P + Q))$. When \mathbf{L} is a sparse matrix or linear mapping, the scaling depends on

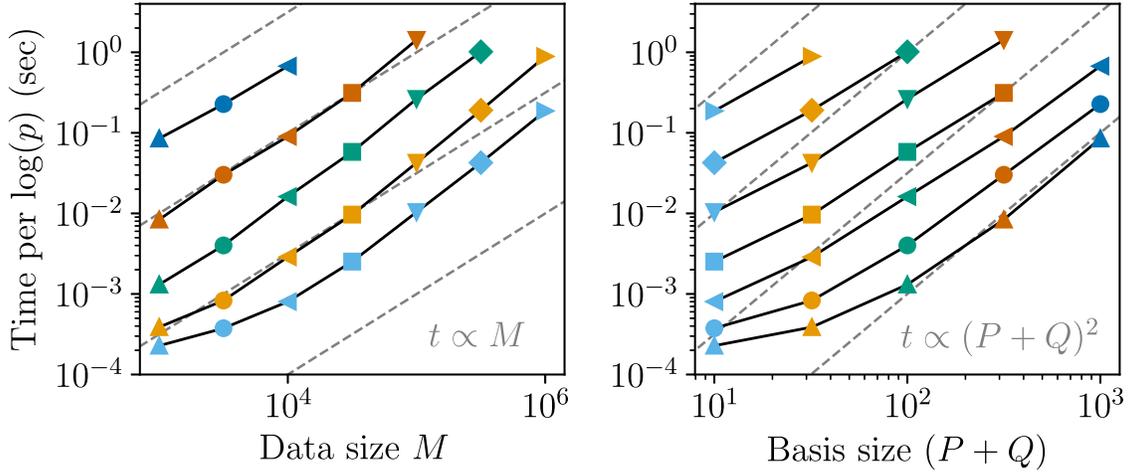


Figure 6. Computation time of the marginal log-likelihood (Equations (10) and (13)) when the data covariance matrix \mathbf{K} is diagonal and \mathbf{L} is the identity mapping as a function of data set size M (left panel) and basis size $P+Q$ (right panel). Values with the same marker shape were computed at the same data set size M . Values with the same marker color were computed at the same data set size $P+Q$. Polynomials of the form given in the bottom right corner of each panel are shown as dashed gray lines.

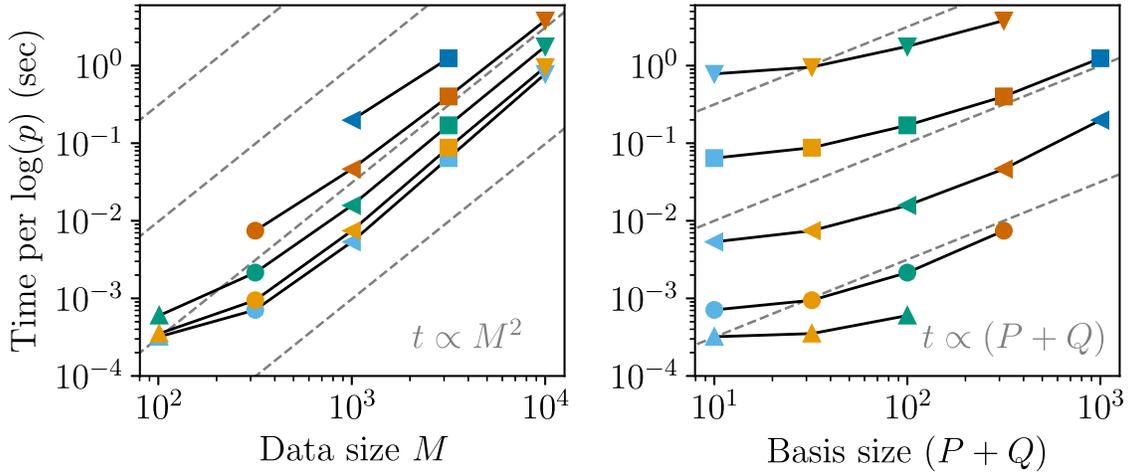


Figure 7. Computation time of the marginal log-likelihood when the data covariance matrix \mathbf{K} is not diagonal and \mathbf{L} is a dense matrix. See caption of Figure 6 for a description of figure elements.

its exact structure. An LSF that varies with wavelength can be represented by a banded matrix, which will be sparse if the spectrum spans many resolution elements. If the bandwidth of \mathbf{L} is independent of the size of the data set, the computation time of this product grows as $\mathcal{O}(M(P+Q))$.

We consider covariance matrices \mathbf{K} that are either diagonal or general. If \mathbf{K} is diagonal, $\mathbf{K}^{-1}(\mathbf{L}\mathbf{B})$ requires exactly $M(P+Q)$ multiplications. When \mathbf{K} is a general covariance matrix, we decompose it into its Cholesky factors and left multiply $\mathbf{L}\mathbf{B}$ by \mathbf{K}^{-1} by solving the linear problem $\mathbf{L}\mathbf{B} = \mathbf{K}\mathbf{X}$. The time needed to factor \mathbf{K} grows as $\mathcal{O}(M^3)$ but only needs to be done once per set of observations. The time needed to solve the linear problem grows as $\mathcal{O}(M^2(P+Q))$.

To empirically confirm these growth rates, we timed how long it takes to evaluate the log-likelihood and its gradient for a range of data set sizes M and basis sizes $P+Q$ and three \mathbf{L} and \mathbf{K} structure scenarios. The scenarios are: \mathbf{L} is the identity mapping, \mathbf{K} is diagonal; \mathbf{L} is a dense matrix, \mathbf{K} is general; and \mathbf{L} is a sparse, banded matrix and \mathbf{K} is diagonal. The first two scenarios are the fastest and slowest combination. The third scenario is more typical for a spectrum; the data uncertainty is

diagonal and the LSF has finite extent. The evaluation time of the log-likelihood as a function of M and $P+Q$ for these three scenarios is shown in Figures 6–8. We do not show the evaluation time of the gradient because it behaves in the same way as the evaluation time of the log-likelihood in all three scenarios; the most expensive step of the two calculations is the same.

The dependence of computation time on M and $P+Q$ generally agrees with the predictions based on the two most time-consuming steps. At low M and in particular at low $P+Q$, the computation time is either overhead-dominated or evenly split between the most time-consuming steps and other steps. When $M \gtrsim 10^5$, computation time increases faster than expected purely from the growth rate of the required number of operations (see, e.g., the left panel of Figure 6). This excess increase in computation time is most likely due to changes in memory bandwidth, as the size of matrix rows and columns increases past the size of the highest-level CPU cache on the laptop used to run these tests.

To put these data set sizes into context, a Sloan Digital Sky Survey Baryon Oscillation Spectroscopic Survey or Apache

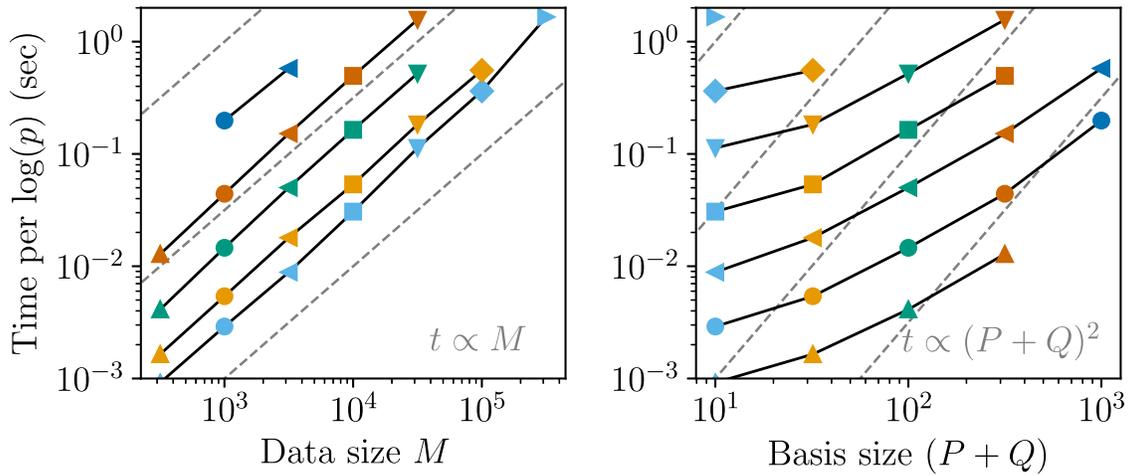


Figure 8. Computation time of the marginal log-likelihood when the data covariance matrix \mathbf{K} is diagonal and \mathbf{L} is a sparse, banded matrix. See the caption of Figure 6 for a description of the figure elements.

Point Observatory Galactic Evolution Experiment spectrum is $\sim 10^3$ pixels long, a *Hubble Space Telescope* Cosmic Origins Spectrograph (*HST*-COS) spectrum is $\sim 10^4$ pixels long, and a spectrum from an echelle spectrograph such as the Ultraviolet and Visual Echelle Spectrograph on the Very Large Telescope or the *Magellan* Inamori Kyocera Echelle spectrograph is $\sim 10^5$ – 10^6 pixels long. The uncertainties associated with these spectra are usually assumed to be diagonal and the LSFs are acceptably described by sparse, banded matrices, so the computation times given in Figures 6 and 8 should apply.

Appendix B

Convergence and Effective Sample Generation Rate of MCMC

In ISM absorption spectra, it is common to have multiple lines in a spectrum with shared parameters. These lines can be from the same species, e.g., the Lyman series, or from different species, e.g., from Mg I, Zn II, and Cr II, which have overlapping lines in the near-ultraviolet. When these lines are in different parts of a spectrum, each part needs its own continuum parameters. This is a case in which analytic marginalization can potentially be more efficient than MCMC marginalization.

We compare how quickly MCMC is done using each of the two methods is post-convergence and how efficient MCMC done using each method is pre-convergence. Which comparison is more informative for choosing a method to use will depend on the purpose of the MCMC run. If the goal of an MCMC run is to estimate some value at low-to-moderate precision, the rate of convergence will be the more important factor. If the goal is instead to estimate some value at high precision, the burn-in period will usually be a small fraction of the total chain and post-convergence efficiency will be more important.

We consider a case where there are N absorption lines with shared central velocities and widths and independent column densities. Each absorption line is in a different spectral region. The continuum in each spectral region is a polynomial of degree M . The marginalized likelihood has $2 + N$ absorption line parameters. The unmarginalized likelihood has $2 + N$ absorption line parameters and $N \times M$ continuum parameters. We use the `emcee` implementation of the Goodman and Weare affine-invariant MCMC ensemble sampler to generate draws

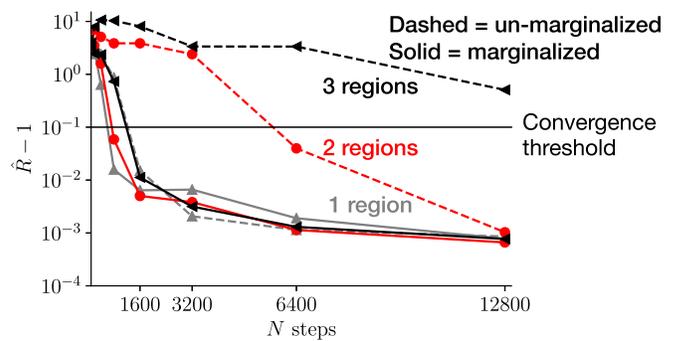


Figure 9. Convergence rate of MCMC with analytic and numerical continuum parameter marginalization for absorption line analysis problems with different complexities. The convergence diagnostic (y-axis) is the Rubin–Gelman statistic, an estimate of how much smaller the Monte Carlo error of an MCMC-based parameter estimate can get. Each line shows the evolution of this convergence diagnostic as a function of the number of MCMC steps taken (x-axis). Line styles indicate whether continuum parameters are marginalized over analytically (solid) or included in MCMC (dashed). Line colors and markers indicate the number of spectral regions being analyzed simultaneously; each region has its own set of continuum parameters. The Rubin–Gelman statistic and the problem setup are discussed in more detail in Appendix B.

from the posterior corresponding to each of these likelihoods. We use the minimum number of walkers, which is twice the number of parameters.

We use the Rubin–Gelman statistic \hat{R} (Gelman & Rubin 1992) to assess convergence. The Rubin–Gelman statistic compares the variance between and within different MCMC instances. If the instances have all converged, these two variances should be approximately equal. We run 10 MCMC instances for 12,800 (per walker) steps and compute the Rubin–Gelman statistic from the second half of subchains of length $2^p \times 100$ for $p = 0, 1, \dots, 7$. \hat{R} is computed separately for each parameter. Following common usage, we consider convergence to be reached when the \hat{R} of all parameters is less than 1.1. We run this test for 1, 2, and 3 regions and absorption lines assuming a continuum of degree 1. The value of the \hat{R} as a function of the (total) number of steps is shown in Figure 9. When there is a single region and line, the MCMC marginalization chain takes twice as many steps as the analytic marginalization chain to converge; when there are two regions, it takes eight times as many steps; when there are three, the

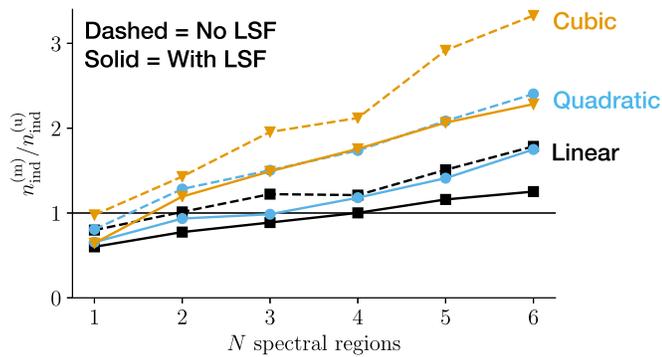


Figure 10. Relative efficiency of MCMC with analytic and numerical continuum parameter marginalization for absorption line analysis problems with different complexities. The relative efficiency is the ratio of the number of independent samples, n_{ind} , generated in the same amount of time by the two marginalization approaches; $n_{\text{ind}}^{(m)}$ uses the analytically marginalized likelihood, $n_{\text{ind}}^{(u)}$ uses the unmarginalized likelihood. The larger the relative efficiency, the more independent samples generated by analytic marginalization. Line colors and markers correspond to different continuum parameterizations: degree 1 polynomial (black squares), degree 2 polynomial (blue circles), and degree 3 polynomial (orange triangles). Line styles indicate whether a nontrivial LSF is used in the analysis. The relative efficiency is shown as a function of the number of spectral regions being analyzed simultaneously; each spectral region has its own set of continuum parameters. The relative efficiency and the problem setup are discussed in more detail in Appendix B.

MCMC marginalization chain has not converged by the maximum chain length of 12,800 while the analytic marginalization chain converges within 1600 steps.

We use the number of independent samples per unit time to assess efficiency. We run MCMC with the marginalized likelihood for 2000 burn-in steps and 8000 converged steps and record the average time per sample, t_s . Because MCMC with the unmarginalized likelihood takes many steps to converge, we use draws from the converged part of the marginalized likelihood chain as a starting point; these draws only have values for the absorption line parameters. At each set of absorption line parameters, we sample a set of continuum parameters from the conditional distribution discussed in Section 2.1. From this starting point, we run MCMC with the unmarginalized likelihood for 4000 burn-in steps and 36,000 converged steps and record the average time per sample. We then compute the average integrated autocorrelation times τ_f of

the walkers in both chains. The number of independent samples per unit time is $n_i = (\tau_f t_s)^{-1}$.

We compute n_i for a number of regions $N = 1, 2, \dots, 6$, continua of polynomial degree $M = 1-3$, and either a trivial LSF or a banded LSF. The ratio $n_{\text{ind}}^{(m)}/n_{\text{ind}}^{(u)}$ for each of these cases is shown in Figure 10. When this ratio is greater than 1, running MCMC with the marginalized likelihood for a fixed amount of time will produce more independent samples than running MCMC with the unmarginalized likelihood for the same amount of time. The greater the number of regions and the degree of the continuum, the greater the efficiency advantage of the marginalized likelihood over the unmarginalized likelihood. This advantage will not depend on the number of data points in each spectral region so long as the LSF is trivial or banded, since in these cases the evaluation time of both likelihoods grows linearly with data set length (see Appendix A.3).

ORCID iDs

Kirill Tchernyshyov  <https://orcid.org/0000-0003-0789-9939>

References

- Buder, S., Asplund, M., Duong, L., et al. 2018, *MNRAS*, **478**, 4513
Casey, A. R. 2016, *ApJS*, **223**, 8
Czekala, I., Andrews, S. M., Mandel, K. S., Hogg, D. W., & Green, G. M. 2015, *ApJ*, **812**, 128
Eilers, A.-C., Hennawi, J. F., & Lee, K.-G. 2017, *ApJ*, **844**, 136
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, **125**, 306
Gelman, A., & Rubin, D. B. 1992, *StaSc*, **7**, 457
Hunter, J. D. 2007, *CSE*, **9**, 90
Jones, E., Oliphant, T., Peterson, P., et al. 2001, SciPy: Open Source Scientific Tools for Python, <http://www.scipy.org>
Liang, C. J., Kravtsov, A. V., & Agertz, O. 2018, *MNRAS*, **479**, 1822
Lindner, R. R., Vera-Ciro, C., Murray, C. E., et al. 2015, *AJ*, **149**, 138
Luger, R., Foreman-Mackey, D., & Hogg, D. W. 2017, *RNAAS*, **1**, 7
Price-Whelan, A. M., Hogg, D. W., Foreman-Mackey, D., & Rix, H.-W. 2017, *ApJ*, **837**, 20
Riener, M., Kainulainen, J., Henshaw, J. D., et al. 2019, *A&A*, **628**, A78
Suzuki, N. 2006, *ApJS*, **163**, 110
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *CSE*, **13**, 22
Wakker, B. P., Hernandez, A. K., French, D. M., et al. 2015, *ApJ*, **814**, 40
Welty, D. E., Lauroesch, J. T., Wong, T., & York, D. G. 2016, *ApJ*, **821**, 118
Woodgate, B. E., Kimble, R. A., Bowers, C. W., et al. 1998, *PASP*, **110**, 1183
Zhu, G., & Ménard, B. 2013, *ApJ*, **770**, 130