Draft version January 4, 2019
Typeset using LaTeX **manuscript** style in AASTeX62

Analytic marginalization of absorption line continua and other multiplicative linear nuisance parameters

Kirill Tchernyshyov[1]

[1] *Department of Physics and Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*

ABSTRACT

Absorption line spectroscopy can be 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 datasets consisting of thousands or even millions of spectra. Simultaneous probabilistic modeling of absorption features and the intrinsic continuum shape is a promising approach for automating this analysis. This requires marginalizing over continuum parameters. Doing this numerically, e.g. using Markov chain Monte Carlo, for large numbers of spectra is impractical. A common way of parametrizing a continuum is as a linear function of such as a polynomial or spline. When such a parametrization is used, it is in fact possible to reduce continuum parameter marginalization to an integral over a multivariate normal distribution, which has a known closed form. We implement this integration in the open-source `python` package `name`. The availability of a closed form makes marginalization over different possible linear continuum functions trivial. To demonstrate the power of parameter and parametrization marginalization, we com-

Corresponding author: Kirill Tchernyshyov
ktcherny@gmail.com

pare the accuracy to which absorption line parameters can be recovered using different continuum placement methods. Parameter and parametrization marginalization are consistently better or as good as other ways of continuum placement and continuum model selection.

Keywords: methods: statistical

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 separate the absorption lines and the intrinsic spectrum, typically referred to as the continuum, of the illuminating background source towards 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 intervening absorption, fitting a function to these regions, and using this function to interpolate over regions that do contain absorption. 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 practically be done for datasets containing thousands or even millions of spectra. The second is that it is difficult to determine the accuracy and precision of an estimator that involves a human. It is therefore not possible to determine whether there are obvious alternative estimators that would use the data more efficiently.

An increasingly popular 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, and `sick`, Casey 2016). In these packages, continuum parameter marginalization is done numerically, using MCMC. The authors of two of these packages point out that including large numbers of continuum parameters in MCMC sampling leads to long convergence and autocorrelation times. To keep the number of continuum parameters low, the

packages either do not support (`BayesVP`) or advise against (`sick`) including continuum parameters when simultaneously analyzing multiple spectral segments.

In these packages and in much of the absorption line analysis literature, the continuum is assumed to be a low order polynomial or spline. While these are non-linear functions of wavelength, they are linear functions of the polynomial or spline coefficients. This linearity means that if some additional assumptions hold, it is possible to marginalize over these coefficients analytically. Analytic marginalization has several advantages over numerical marginalization. First, it can speed up MCMC-based inference for absorption line parameters by reducing the dimensionality of the problem. Second, because the gradient of the continuum parameter-marginalized likelihood is also available in closed form, it is possible to efficiently optimize for absorption line parameters while keeping the robustness provided by marginalizing out nuisance parameters. The gradient can also be used in gradient-aware versions of MCMC such as Hamiltonian Monte Carlo (Duane et al. 1987). Finally, it makes marginalization over different possible continuum parametrizations computationally trivial—simply add continuum-marginalized likelihoods that assume different parametrizations together. Marginalization over parametrizations allows a greater degree of automation and systematization of absorption line inference.

The assumptions required by the method are: that the continuum can be expressed as a linear function (not necessarily a polynomial); that the prior on the parameters of this linear function are either improper uniform or Gaussian; and that residuals from the model are Gaussian. If these assumptions hold, then the posterior probability distribution function of the continuum parameters given a model for the absorption is a multivariate Gaussian. Marginalizing over this multivariate Gaussian just modifies the covariance matrix of the multivariate Gaussian distribution describing the residuals. When a set of absorption line parameters is specified, the continuum parameters can be treated as additive, rather than multiplicative, linear nuisance parameters. An explanation of marginalization of additive linear nuisance parameters in an astronomical context is given in Luger et al. (2017). This approach to marginalizing over multiplicative linear nuisance parameters has been

used, for example, to analyze sparsely sampled radial velocity measurements (Price-Whelan et al. 2017).

Models for absorption line spectra have features, such as the presence of a line spread function (LSF), which should be accounted for to more efficiently compute marginalized likelihoods and likelihood gradients. In this work, I derive expressions for the likelihood and its gradient that account for these features. This derivation is given in Section 2. I have created a package, `name`[1], for evaluating these expressions. Appendix A describes the package and gives some performance benchmarks. The performance of continuum parameter and parametrization marginalization is explored in Section 3. I discuss strengths and weaknesses of this method in Section 4 and conclude in Section 5.

2. ASSUMPTIONS AND FORMALISM

We assume the following model for a data vector $\mathbf{y}$ of length M:

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

$\mathbf{L}$ is a linear mapping from R^N to R^M. If $\mathbf{y}$ is a spectrum, $\mathbf{L}$ would be the instrumental line spread function. If we were fitting an absorption Voigt profile to this spectrum, $\mathbf{d}(\theta)$ would be the transmittance as a function of wavelength and θ would include the center, Doppler width, and total optical depth of the Voigt profile. The length of the vector $\mathbf{d}(\theta)$ is N. The m_i are the MLNPs, the $\mathbf{a}_{m,i}$ are the multiplicative basis elements, and $\boldsymbol{\mu}_m(\theta)$ is the mean of the multiplicative linear part of the model. These parameters and basis elements would be a model for the continuum emitted behind the absorbing material. The b_i are additive linear nuisance parameters with basis $\mathbf{a}_{b,i}$ and $\boldsymbol{\mu}_b(\theta)$ is the mean of the additive linear part of the model. These parameters would be a model for any emission happening between the observer and the absorbing material, such as sky line emission. $\boldsymbol{\varepsilon}$ is the residual vector, which we assume to be normally distributed with mean zero and covariance matrix $\mathbf{K}$:

$$p(\boldsymbol{\varepsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{K}). \tag{2}$$

[1] available HERETKTK

Collecting the multiplicative and additive basis elements $\mathbf{a}_{m,i}$ and $\mathbf{a}_{b,i}$ into matrices $\mathbf{A_m}$ and $\mathbf{A_b}$ and converting the vector $\mathbf{d}(\theta)$ into the diagonal matrix $\mathbf{D}_\theta \equiv \mathrm{diag}\,(\mathbf{d}(\theta))$,

$$\mathbf{y} = \mathbf{L}\left(\boldsymbol{\mu}_b(\theta) + \mathbf{A_b}\mathbf{b} + \mathbf{D}_\theta\left(\boldsymbol{\mu}_m(\theta) + \mathbf{A_m}\mathbf{m}\right)\right) + \boldsymbol{\varepsilon} \tag{3}$$

$$\equiv \mathbf{L}\left(\boldsymbol{\mu}_b(\theta) + \mathbf{D}_\theta\boldsymbol{\mu}_m(\theta) + \mathbf{B}\mathbf{c}\right) + \boldsymbol{\varepsilon}. \tag{4}$$

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

$$\mathbf{B} = \begin{bmatrix}\mathbf{D}_\theta\mathbf{A_m} & \mathbf{A_b}\end{bmatrix} \qquad\qquad \mathbf{c} = \begin{bmatrix}\mathbf{m}\\ \mathbf{b}\end{bmatrix}. \tag{5}$$

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

$$p_n(\mathbf{c}) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}\right) \text{ (normal)} \qquad \text{and} \qquad p_u(\mathbf{c}) = \prod_{i=1}^{P+Q} Z_i^{-1} \text{ (uniform)}, \tag{6}$$

where Z_i can be any positive real number.

2.1. *Conditional probability of the nuisance parameters*

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

$$\hat{\mathbf{c}}_{n/u} = \mathbf{C}_{n/u}^{-1}\mathbf{B}^T\mathbf{L}^T\mathbf{K}^{-1}\mathbf{r}, \tag{7}$$

where $\mathbf{r}$ is the vector of residuals

$$\mathbf{r} = \mathbf{y} - \mathbf{L}\left(\boldsymbol{\mu}_b(\theta) + \mathbf{D}_\theta\boldsymbol{\mu}_m(\theta)\right) \tag{8}$$

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

$$\mathbf{C}_n = \boldsymbol{\Lambda}^{-1} + \mathbf{B}^T\mathbf{L}^T\mathbf{K}^{-1}\mathbf{L}\mathbf{B} \tag{9}$$

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

$$\mathbf{C}_u = \mathbf{B}^T\mathbf{L}^T\mathbf{K}^{-1}\mathbf{L}\mathbf{B} \tag{10}$$

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 $\mathbf{LB}\hat{\mathbf{c}}$ 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.

2.2. *Marginal likelihood*

Assuming the proper prior $p_n(\mathbf{c})$, marginalizing over $\mathbf{c}$ following e.g. Luger et al. (2017) or Rasmussen & Williams (2006) gives

$$p_n(\mathbf{y}|\theta, \mathbf{L}, \mathbf{B}, \mathbf{K}, \boldsymbol{\Lambda}) = \int_{-\infty}^{+\infty} p(\mathbf{y}|\mathbf{c}, \theta, \mathbf{L}, \mathbf{B}, \mathbf{K}, \boldsymbol{\Lambda}) p_n(\mathbf{c}) \, \mathrm{d}\mathbf{c} \tag{11}$$

$$= (2\pi)^{-\frac{M}{2}} \det(\mathbf{K})^{-\frac{1}{2}} \det(\boldsymbol{\Lambda})^{-\frac{1}{2}} \det(\mathbf{C}_n)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{r}^T \mathbf{K}^{-1} (\mathbf{r} - \hat{\mathbf{r}}_n)\right], \tag{12}$$

where

$$\hat{\mathbf{r}}_{n/u} = \mathbf{LB}\hat{\mathbf{c}}_{n/u}. \tag{13}$$

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

$$p_u(\mathbf{y}|\theta, \mathbf{L}, \mathbf{B}, \mathbf{K}) = \int_{-\infty}^{+\infty} p(\mathbf{y}|\mathbf{c}, \theta, \mathbf{L}, \mathbf{B}, \mathbf{K}) p_u(\mathbf{c}) \, \mathrm{d}\mathbf{c} \tag{14}$$

$$= \left(\prod_{i=1}^{P+Q} Z_i^{-1}\right) (2\pi)^{-\frac{M-(P+Q)}{2}} \det(\mathbf{K})^{-\frac{1}{2}} \det(\mathbf{C}_u)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{r}^T \mathbf{K}^{-1} (\mathbf{r} - \hat{\mathbf{r}}_u)\right]. \tag{15}$$

The marginal likelihood p_u will be proper whenever $\mathbf{C}_u$ is positive definite, which will be the case whenever $\mathbf{LB}$ is 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 $\boldsymbol{\Lambda}^{-1}$ is always positive definite and $\mathbf{B}^T\mathbf{L}^T\mathbf{K}^{-1}\mathbf{LB}$ is always at least positive semi-definite.

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

$$\begin{aligned}\nabla \log(p)(\mathbf{d}(\theta)) = &\left(\mathbf{L}^T\mathbf{K}^{-1}\left(\mathbf{r} - \hat{\mathbf{r}}_{n/u}\right)\right) \odot \left(\mathbf{B}'\hat{\mathbf{c}} + \boldsymbol{\mu}_m\right) \\ &- \frac{1}{2}\left(\left(\mathbf{C}_{n/u}^{-1}\mathbf{B}'^T\right) \odot \left(\mathbf{B}^T\mathbf{L}^T\mathbf{K}^{-1}\mathbf{L}\right) + \left(\mathbf{C}_{n/u}^{-1}\mathbf{B}^T\mathbf{L}^T\mathbf{K}^{-1}\mathbf{L}\right) \odot \mathbf{B}'^T\right)\mathbf{1},\end{aligned} \tag{16}$$

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 $\mathbf{d}(\theta)$:

$$\mathbf{B}' = \sum_{i=1}^{N} \frac{\partial \mathbf{B}}{\partial d_i(\theta)} \tag{17}$$

$$= \sum_{i=1}^{N} \left[\mathbf{J}^{i,i}\mathbf{A_m} \;\; 0 \times \mathbf{A_b}\right] \tag{18}$$

$$= \left[\mathbf{A_m} \;\; \mathbf{0}\right], \tag{19}$$

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 16 is the gradient of the argument of the exponentials in Equations 11 and 14. The second row is the gradient of $\log\left(\det\left(\mathbf{C}_{n/u}\right)\right)$.

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

$$\nabla \log(p)(\boldsymbol{\mu}_m(\theta)) = \mathbf{D}_\theta\mathbf{L}^T\mathbf{K}^{-1}\left(\mathbf{r} - \hat{\mathbf{r}}_{n/u}\right) \tag{20}$$

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

$$\nabla \log(p)(\boldsymbol{\mu}_b(\theta)) = \mathbf{L}^T\mathbf{K}^{-1}\left(\mathbf{r} - \hat{\mathbf{r}}_{n/u}\right) \tag{21}$$

3. PRACTICAL TEST CASES

Is marginalization of continuum parameters, analytic or not, actually useful? We consider two metrics: how marginalizing over, instead of fitting for, continuum parameters affects the error with which absorption line parameters can be measured and how marginalizing analytically instead of numerically affects the speed of MCMC-based inference.

For the error metric, we consider two possible cases: one in which the continuum parametrization is known and one in which it is necessary to choose a continuum parametrization from a set of possibilites. Our suggested method is (analytically) marginalizing over continuum parameters, in the first case, and marginalizing over continuum parameters and parametrizations, in the second. When the parametrization is known, all methods we consider are equal when SNR is sufficiently high but marginalization is consistently more robust at low SNR. When the parametrization is not known, marginalization over parametrizations has the lowest over-all error rate among the methods we consider. Furthermore, its error rate is close to the error rate of parameter marginalization with a known parametrization. These two cases are examined in Sections 3.1 and 3.2.

We examine how two speed metrics change as the complexity of an inference problem increases: the number of iterations required for MCMC to converge and the number of independent samples generated per unit time. The basic problem is analyzing a single line on a continuum whose parametrization is known. To build up more complicated problems, we add more spectral segments each of which has its own continuum and contains another absorption line. All of these absortion lines share widths and central velocities but have independent column densities. Problems with this structure arise when analyzing multiple lines from a single species or from multiple species that can be assumed to share component structure. The convergence speedup from using analytic marginalization is dramatic, reaching a full order of magnitude difference in the number of required iterations with as few as three spectral segments. Analytic marginalization yields more independent samples per unit time when there are multiple spectral segments with high-order continua. For example, when there are six spectral segments with 3rd order continua, analytic marginalization is three time faster than numerical marginalization. When there are few spectral segments, analytic marginalization is slightly slower or of comparable speed to numerical marginalization. These metrics are examined in Section 3.3.

3.1. *Marginalization over parameters*

We consider the problem of measuring the column density of a single well-resolved, unsaturated absorption line superimposed on a continuum whose parametrization is known but whose parameters

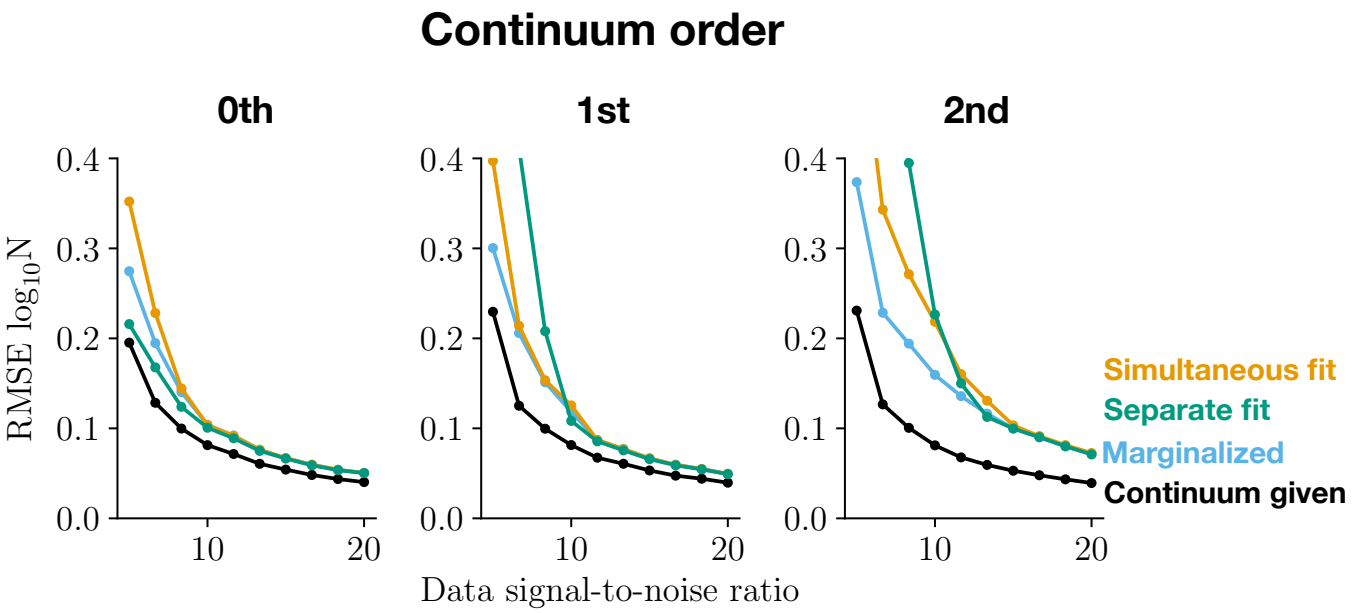

Figure 1. Accuracy and precision of different methods of measuring the column density of a single line superimposed on a continuum with known parametrization. The accuracy/precision is defined in terms of the root mean square error (RMSE) of the logarithm of column density measurements. The signal-to-noise ratios (SNRs) of the artifically generated spectra used for this test are shown on the x-axis of each panel. The panels correspond to different continuum parametrizations, from left to right: 0th order polynomial, 1st order polynomial, 2nd order polynomial. The line colors indicate different measurement methods, which are listed in the figure legend. These methods are explained in detail in Section 3.1.

are not known. To do this, we generate spectra containing a fixed absorption line with different continuum parametrizations and signal-to-noise ratios (SNRs) and measure the column density from each of these spectra. The continuum parametrizations we use are polynomials of order 0, 1, and 2. We consider SNRs between 5 and 25. The summary statistic that we use to indicate quality is the root mean square error (RMSE) of the base ten logarithm of the column density ($\log_{10} \mathrm{N}$). To compute each RMSE, we generate 1000 spectra. We use the logarithm of the column density because physical constants (e.g. oscillator strengths) cancel in that RMSE calculation.

We measure the column density in four ways: (1) supply the correct continuum parameters and only fit for the absorption line parameters; (2) simultaneously fit for continuum and absorption line parameters; (3) analytically marginalize over continuum parameters and fit for absorption line parameters; and (4) use the absorption line parameters recovered using method (1) to define a

line-free spectral region, fit continuum parameters just to this region, and with those continuum parameters fit for the absorption line parameters. The first method is meant to set a lower limit on the RMSE of $\log_{10} \mathrm{N}$ as a function of SNR. The second and third methods are two possible ways of automatically modeling the continuum. The fourth method is meant to approximate the actions of a human manually analyzing a spectrum. We assume the human can correctly estimate the continuum by eye, correctly estimate the best-fit absorption line profile by eye given this continuum, and finally use this profile to determine which part of the spectrum is not affected by the line.

The RMSEs obtained using these different methods are shown in Figure 1. Above an SNR of 10-15, all methods where the correct continuum parameters are not known a priori are equal. At and below that SNR range, marginalization has a lower RMSE than both simultaneous fitting and the human-like analysis. The advantage of marginalization over the other methods becomes greater as the continuum parametrization becomes more complex.

3.2. *Marginalization over parametrizations*

Next, we consider a problem where there is still a single well-resolved and unsaturated absorption line but where we only know that the continuum belongs to a *family* of possible continuum parametrizations. As in the previous section, we consider three possible continuum parametrizations: 0th, 1st, and 2nd order polynomials. The approach, simulating spectra, measuring $\log_{10} N$ for each simulated spectrum, and computing the RMSE of $\log_{10} N$, is also the same. However, we consider a different range in SNR: 10 to 100.

We measure the column density in three ways: (1) supply the correct parametrization and marginalize over its parameters; (2) assume the most complicated of the three parametrizations and marginalize over the parameters; (3) repeatedly use the likelihood ratio test to choose a parametrization and fit for parameters; (4) marginalize over parametrizations as well as parameters. Method (1) is meant to establish a reference minimum RMSE for this test case. Method (2) is a conservative assumption that can be made when the family of possible parametrizations is nested—a polynomial of order n with leading coefficient 0 is a polynomial of order $n - 1$. Methods (3) and (4) are different ways of automatically accounting for the different possible parametrizations, in one case (3) by selecting a

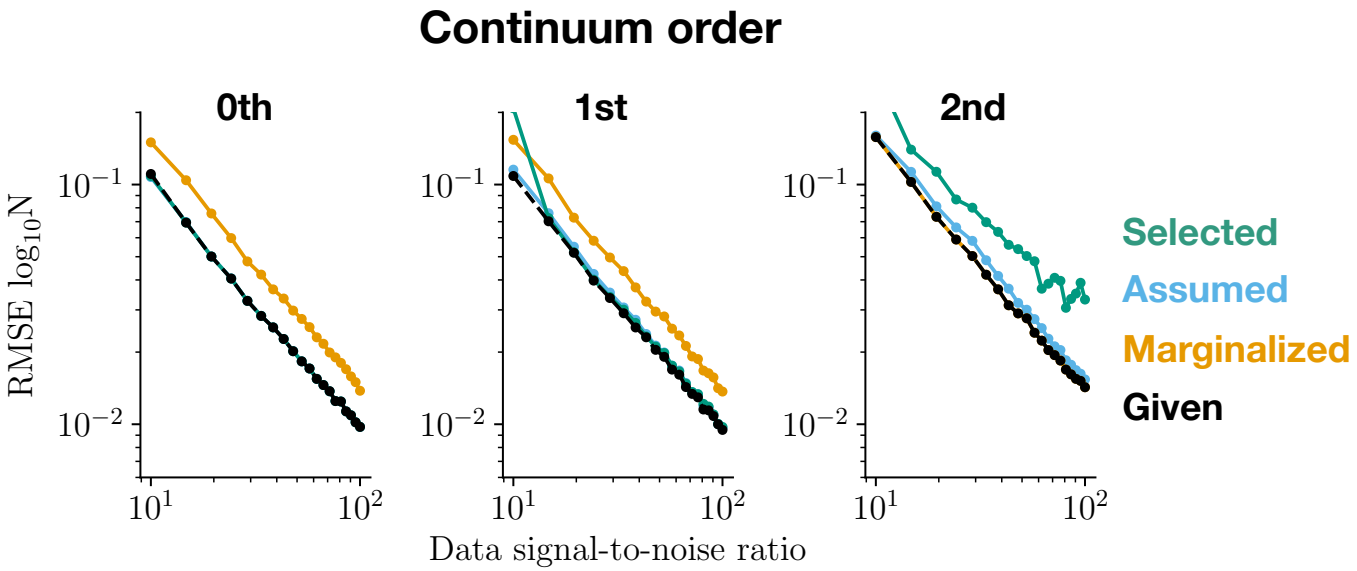

Figure 2. Accuracy and precision of different methods of measuring the column density of a single line superimposed on a continuum with unknown parametrization. The accuracy/precision is defined in terms of the root mean square error (RMSE) of the logarithm of column density measurements. The signal-to-noise ratios (SNRs) of the artifically generated spectra used for this test are shown on the x-axis of each panel. The panels correspond to different true continuum parametrizations, from left to right: 0th order polynomial, 1st order polynomial, 2nd order polynomial. The line colors indicate different measurement methods, which are listed in the figure legend. These methods are explained in detail in Section 3.2.

parametrization and in other (4) by averaging over parametrizations. The likelihood function that we maximize when using method (4) is the weighted sum of the continuum-marginalized likelihoods of the three possible models. The weights are the prior probabilities of each of the models; we assume all three are equally likely.

The RMSEs of the four methods are shown in Figure 2. Three ways of interpreting results. Robustness of solution to decrease in SNR and shortening of spectrum relative to number of parameters; RMSE obtainable by different methods at fixed SNR; SNR required by different methods to obtain the same SNR. By robustness, we mean that RMSE has a consistent scaling with SNR and true continuum order. The estimator does not break down as SNR decreases below some breakdown point or as the number of continuum parameters increases. The reference, conservative, and parametrization-marginalization methods are robust for SNRs between 10 and 100. On the other

hand, the parametrization-selection method is not robust; its RMSE blows up at an SNR of 10 for spectra with 1st order continua and at all SNRs considered for 2nd order continua. Parametrization selection should not be used for noisy spectra or spectra with a high ratio of parameters to data points.

In terms of RMSE, the parametrization-marginalized estimator is nearly as good as the reference estimator. The ratio RMSEm/RMSEr of the parametrization-marginalized RMSE, RMSEm, to the reference RMSE, RMSEr, is 1, 1.04, and 1.1 for spectra with 0th, 1st, and 2nd order continua. For spectra with 0th or 1st order continua, the conservative estimator is significantly worse than the reference estimator. RMSEc/RMSEr is 1.5, 1.5, and 1, respectively, where RMSEc is the RMSE of the conservative estimator. These ratios are approximately constant across the entire SNR range considered. When analyzing an already acquired sample of observations in which a variety of continuum parametrizations are present, using the parametrization-marginalization estimator rather than the conservative estimator will, on average, yield higher accuracy and precision.

To get a sense of how much better the parametrization-marginalization estimator is than the conservative estimator, we can look at the SNR the two estimators require to achieve the same RMSE. For spectra with 0th, 1st, and 2nd continua, the ratio of the required SNRs SNRc/SNRm is 1.6, 1.5, and 0.9. These ratios are consistent across the entire considered SNR range. Assuming that SNR is proportional to the square root of integration time, as is the case for Poisson noise-limited observations, these SNR ratios can be converted to required observing time ratios. Reaching the RMSE of the parametrization-marginalized estimator with the conservative estimator takes 1.26, 1.22, and 0.95 times as much observing time. When designing an observing strategy to meet a column density RMSE requirement, using the parametrization-marginalized estimator rather than the conservative estimator can save observing time given a fixed sample or increase the size of a sample given a fixed amount of observing time.

3.3. *MCMC efficiency*

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

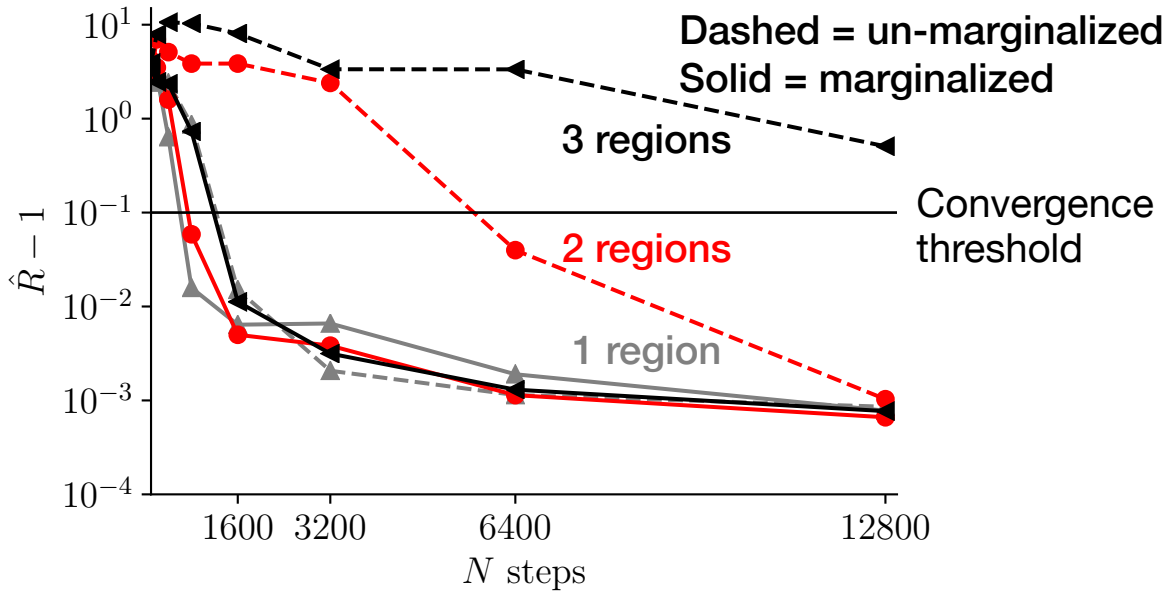

Figure 3. Convergence rate of MCMC with analytic and numerical continuum parameter marginalization for absorption line analysis problems with different complexities. The convergence diagnostic (y-axis) is the Rubin-Gelman statistic, an estimate of how much smaller the Monte Carlo error of an MCMC-based 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 indicates whether continuum parameters are marginalized over analytically (solid) or included in MCMC (dashed). Line colors and markers indicate the number of spectral regions, each of which has its own set of continuum parameters, are being analyzed simultaneously. The Rubin-Gelman statistic and the problem setup are discussed in more detail in Section 3.3.

MgI, ZnII, and CrII in the near ultraviolet. When these lines are in different regions in a spectrum, each region needs its own continuum parameters; this is the case in which BayesVP does not allow the inclusion of continuum parameters in inference. This is one of the scenarios in which analytic marginalization can be more efficient than MCMC marginalization.

We compare the two methods on how quickly MCMC done using each converges and how efficient MCMC done using each is post-convergence. Which comparison is more informative for choosing which 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 to estimate some value at high precision, the burn-in period will usually be a small fraction of the total chain and post-convergency efficiency will be more important.

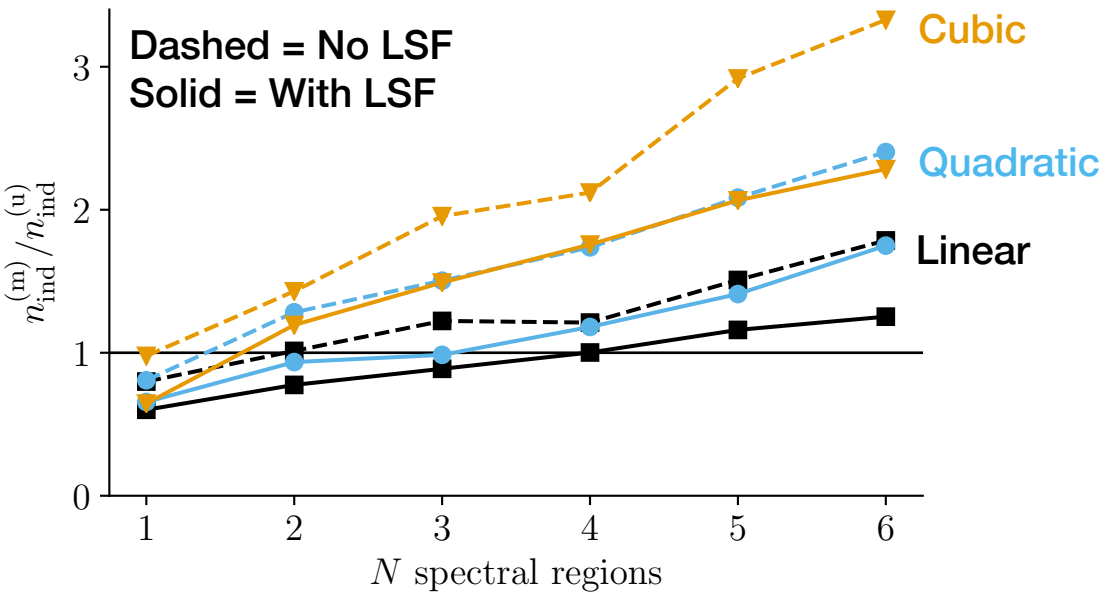

Figure 4. Relative efficiency of MCMC with analytic and numerical continuum parameter marginalization for absorption line analysis problems with different complexities. The relative efficiency is the ratio of the number of independent samples, $n_{\rm ind}$, generated in the same amount of time by the two marginalization approaches; $n_{\rm ind}^{(}m)$ uses the analytically marginalized likelihood, $n_{\rm ind}^{(}u)$ 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 parametrizations: 1st order polynomial (black squares), 2nd order polynomial (blue circles), 3rd order polynomial (orange triangles). Line styles indicate whether a non-trivial line spread function (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 Section 3.3.

We consider a case where there are N absorption lines with shared velocity structure, i.e. central velocity and velocity width, but with different amplitudes. Each absorption line is in a different spectral region. The continuum in each spectral region is a polynomial of order 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 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 $\hat{R}$ CITEP to assess convergence. - TKTK WHAT IS THE RUBIN-GELMAN STATISTIC We run ten MCMC instances for 12800 (per-walker) steps and compute the Rubin-

Gelman statistic from the second half of sub-chains of length $2^p \times 100$ for $p = 0, 1, \ldots, 7$. $\hat{R}$ 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 1.01. We run this test for 1, 2, and 3 regions and absorption lines assuming a continuum of order 1, i.e. a straight line. The value of the $\hat{R}$ as a function of (total) number of steps is shown in Figure 3. 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 MCMC marginalization chain has not converged by the maximum chain length of 12800 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 36000 converged steps and record the average (wall) 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 $N = 1, 2, \ldots, 6$, continua of polynomial order $M = 1$, 2, and 3, and either a trivial LSF or a banded LSF. The ratio $n_i^{\text{marg}}/n_i^{\text{unmarg}}$ for each of these cases is shown in Figure 4. 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 order 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 datapoints in each spectral region so long as the LSF is trivial or banded, since the evaluation time of both likelihoods grows linearly with dataset length (see Section A.3).

4. DISCUSSION

4.1. *Assumptions and consequences*

The explicit assumptions of our analytic marginalization method are that the continuum is a linear function, that the coefficients of this linear function are unconstrained, that uncertainties in the observations are (possibly multivariate) Gaussian, and that the covariance matrix of this Gaussian does not depend on the continuum. These assumptions obviously do not strictly hold for any dataset. For example, unconstrained

coefficients never hold because no background source produces negative flux; this is largely irrelevant in practice. A more relevant example is data in the low photon count regime, which are better described by a Poisson distribution than a Gaussian distribution. This is particularly important when the uncertainties on the measurements are themselves highly uncertain and should be explicitly modeled. In that case, the uncertainties will depend on the Poisson intensity function, which explicitly depends on the continuum. Analytic marginalization of the kind described in this work should not be applied to low SNR X-ray or UV spectra.

An implicit assumption of our method is that the absorption model is realistic. For analytic marginalization to be useful, it must be possible for the absorption model to correctly describe the actually present absorption features. 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 un-modeled line will bias the continuum model. In short, improvements in continuum modeling cannot solve problems of absorption model misspecification.

The continuum models envisaged in this work will usually be effective descriptions rather than (often non-linear) physical descriptions. Most continua whose variation is over longer wavelength scales than the width of absorption lines in question can be approximated in this way. Examples of background sources whose continua can be accurately described in this way include quasars and (particularly rapidly rotating) hot stars. With flexible linear models such as splines, it is possible to describe somewhat complicated pseudo-continua such as stellar wind lines. For even more complicated pseudo-continua such as those of cool stars (Zasowski et al. 2015, e.g.), it is necessary to use a non-linear 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 non-linear model.

4.2. *Interpreting the test cases/the possibilities opened up by analytic marginalization*

The test cases in Section 3 showed that marginalization over continuum parameters and parametrizations is more precise, accurate, and robust than the alternatives. - the exact amount of goodness in that section is, of course, dependent on the line parameters, the resolution, and so on; the general trend does not. On its own, analytic marginalization is just a potentially more computationally efficient way of implementing an existing inference approach. It also allows two qualitatively new approaches: continuum model averaging and absorption parameter optimization with a continuum-marginalized likelihood function.

The test case in Section 3.2 combines both of these approaches—optimizing an absorption parameter likelihood function where the parametrization and parameters of the continuum have been marginalized over. Analytic marginalization makes this possible in two ways: availability of closed form likelihoods and availability of gradients of closed form likelihoods. This is useful for dealing with large surveys. Analyses

of absorption lines in tens of thousands of spectra (Zhu & Ménard 2013; Zasowski et al. 2015, e.g.) cannot practically be done with MCMC. With analytic marginalization, it is possible to at least marginalize over continuum parameters. The results of the test cases suggest that this approach could mean a non-trivial improvement in the accuracy and precision of absorption line measurements.

In cases where MCMC is possible, combining continuum parametrization marginalization with a probabilistic specification of absorption component structure would allow absorption line analysis with human intervention only at the level of specifying priors and candidate continuum parametrizations. Component structure specification can be done in a trans-dimensional inference framework, 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. This way of doing absorption line analysis has two potential advantages. Because it includes marginalization over many different nuisance parameters, it should be pretty robust (e.g. to unresolved saturated structure). Because it is essentially automatic, it allows blinding, which is good for hypothesis testing, and improves reproducibility.

5. CONCLUSION

Absorption lines are an important source of information about stars and the ISM. As larger spectroscopic datasets become available and as reproducibility becomes more standard in astronomy, it becomes necessary to move beyond ad-hoc analysis methods, particularly ones in which a human directly interacts with the spectra. 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 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, I have shown that in many cases, it is possible to replace this numerical marginalization with analytic marginalization (Section 2). Analytic marginalization speeds up MCMC-based analyses in problems with many continuum parameters (Section 3.3). The continuum parameter-marginalized likelihood can also be used for optimization over absorption line parameters. This approach combines the speed of optimization with the advantages of continuum marginalization. Analytic marginalization over continuum parameters makes it trivial to also marginalize over continuum *parametrizations*. As with parameter marginalizaton, parametrization marginalization can be combined with optimization over absorption line parameters. Parametrization marginalization again reduces the amount of direct human interference in the analysis of individual spectra and will be especially useful in analyses of datasets containing spectra with different continuum shapes.

I have also confirmed that marginalization over continuum parameters and parametrizations indeed improves the accuracy of absorption line parameter measurements. The advantage of parameter marignalization is only significant at low SNRs (Section 3.1). Parametrization marginalization, on the other hand, is significantly more accurate than alternative methods of deciding on a continuum parametrization at all SNRs (Section 3.2).

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

People: Josh Peek, Andrew Fox, Yong Zheng, Andrew Casey, Cameron Liang

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 `name`is implemented (Section A.1), list some of its capabilities (Section A.2), and show how the computation time of different calculations grows with dataset and continuum model size (Section A.3).

A.1. *Implementation*

We have implemented `name`as a pure-Python package with `numpy` and `scipy` as dependencies. `name`does not contain functionality for building LSFs or synthesizing absorption models 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 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 re-used during repeated marginalized likelihood evaluations. In particular, it stores the data covariance matrix $\mathbf{K}$; the $\mathbf{c}$ prior covariance matrix $\mathbf{\Lambda}$ and its explicit inverse, if applicable; and the line spread function-like linear mapping $\mathbf{L}$ and its transpose.

Both covariance matrices can be diagonal or fully general. To ensure a common interface, the package includes the `CovarianceMatrix` class, which defines an interface that ensures necessary calculations can be done, and two subclasses, `DiagonalCovarianceMatrix` and `GeneralCovarianceMatrix`. `GeneralCovarianceMatrix` uses the Cholesky decomposition of the supplied covariance matrix for determinant calculations and left multiplication of matrices and vectors by the inverse of the supplied covariance matrix. 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 linear 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 dataset 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 $\mathbf{LB}$ and $\mathbf{K}^{-1}(\mathbf{LB})$. The amount of time required to compute these products depends on the structure $\mathbf{L}$ 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 $\mathbf{LB}$ 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 its exact structure. One case that is relevant to the analysis of spectra is a $\mathbf{L}$ that represents a line spread function. A line spread function 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 dataset, 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, $\mathbf{K}^{-1}(\mathbf{LB})$ requires exactly $M(P+Q)$ multiplications. When $\mathbf{K}$ is a general covariance matrix, we decompose it into its Cholesky factors and left-multiply $\mathbf{LB}$ by $\mathbf{K}^{-1}$ by solving the linear problem $\mathbf{LB} = \mathbf{KX}$. The time needed to factor $\mathbf{K}$ grows as $\mathcal{O}\left(M^3\right)$ but only needs to be done once per set of observations. The time needed to solve the linear problem grows as $\mathcal{O}\left(M^2(P+Q)\right)$.

To empirically confirm these growth rates, we timed how long it takes to evaluate the log-likelihood and its gradient for a range of dataset sizes M and basis sizes $P+Q$ and three $\mathbf{L}$ and $\mathbf{K}$ structure scenarios. The

scenarios are: **L** is the identity mapping, **K** is diagonal; **L** is a dense matrix, **K** is general; and **L** is a sparse, banded matrix and **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, the line spread function 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 5, 6, and 7. 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 5). This excess increase is 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 dataset sizes into context, a Sloan Digital Sky Survey (SDSS) BOSS or APOGEE spectrum contains $\sim 10^3$ pixels, a Hubble Space Telescope Cosmic Origins Spectrograph (HST-COS) spectrum contains $\sim 10^4$ pixels, 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 contains $\sim 10^5 - 10^6$ pixels. The uncertainties associated with these spectra are usually assumed to be diagonal and the line spread functions are acceptably described by sparse, banded matrices, so the computation times given in Figure 5 and 7 should apply.

REFERENCES

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

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. 1987, PhLB, 195, 216 . http://www.sciencedirect.com/science/article/pii/037026938791197X

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PUBL ASTRON SOC PAC, 125, 306

Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90

Jones, E., Oliphant, T., Peterson, P., et al. 2001, SciPy: Open source scientific tools for Python, , . http://www.scipy.org/

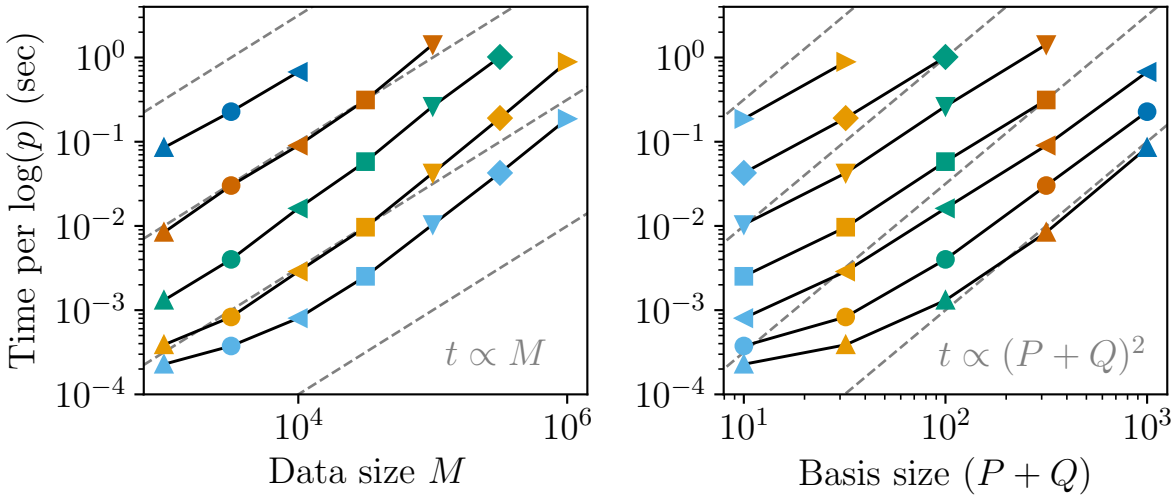

Figure 5. Computation time of the marginal log-likelihood (Equations 11 and 14) when the data covariance matrix $\mathbf{K}$ is diagonal and $\mathbf{L}$ is the identity mapping as a function of dataset size M (left panel) and basis size $P+Q$ (right panel). Values with the same marker shape were computed at the same dataset size M. Values with the same marker color were computed at the same dataset size $P+Q$. Polynomials of the form given in the bottom right corner of each panel are shown as dashed gray lines.

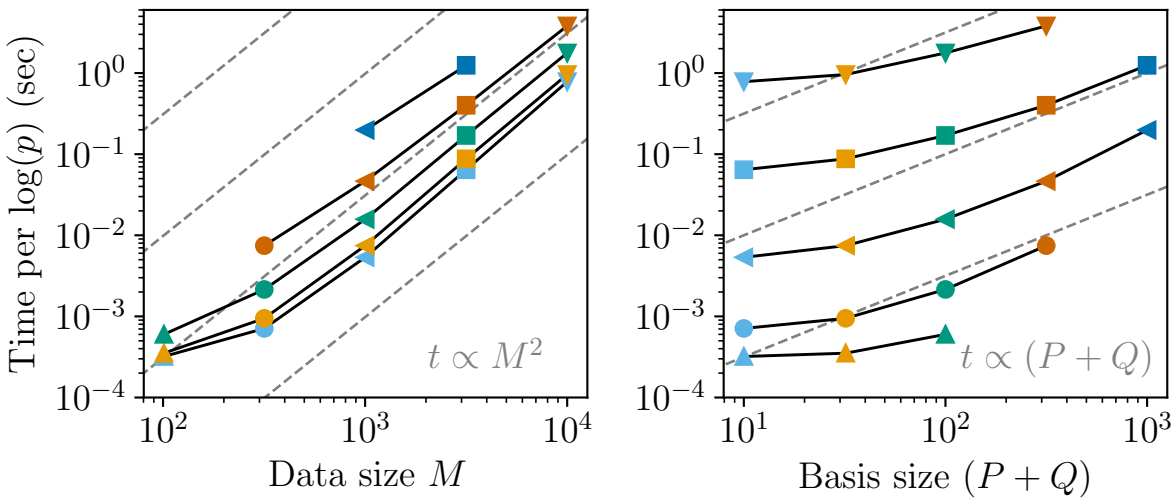

Figure 6. Computation time of the marginal log-likelihood when the data covariance matrix $\mathbf{K}$ is not diagonal and $\mathbf{L}$ is a dense matrix. See caption of Figure 5 for a description of figure elements.

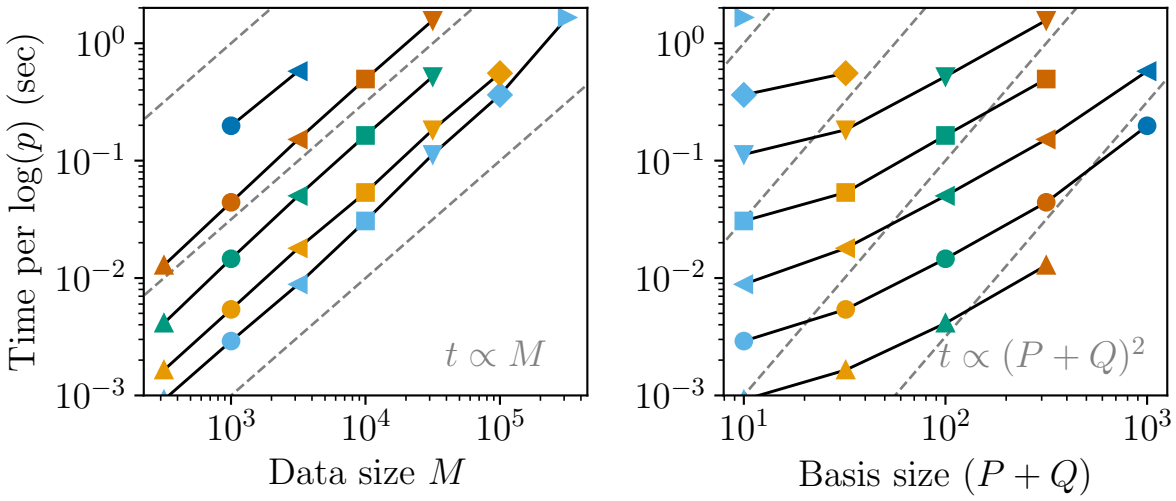

Figure 7. Computation time of the marginal log-likelihood when the data covariance matrix **K** is diagonal and **L** is a sparse, banded matrix. See caption of Figure 5 for a description of figure elements.

Liang, C. J., Kravtsov, A. V., & Agertz, O. 2018, Monthly Notices of the Royal Astronomical Society, 479, 1822

Luger, R., Foreman-Mackey, D., & Hogg, D. W. 2017, Research Notes of the American Astronomical Society, 1, 7

Price-Whelan, A. M., Hogg, D. W., Foreman-Mackey, D., & Rix, H.-W. 2017, ApJ, 837, 20

Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian Processes for Machine Learning (Mit Press)

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Comput. Sci. Eng., 13, 22

Zasowski, G., Ménard, B., Bizyaev, D., et al. 2015, ApJ, 798, 35

Zhu, G., & Ménard, B. 2013, ApJ, 770, 130