Draft version December 19, 2018
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 fitting; speeding it up; well-motivated ways of selecting parameters; large spectroscopic surveys

Keywords: methods: statistical

1. INTRODUCTION

The formation and measurement of many astronomical observables involves processes that combine multiplicatively rather than additively. Light from a distant source is attenuated by intervening interstellar matter (ISM) on its way to a detector; the output of the detector has to be converted to physically meaningful units by applying calibration factors. Often, only some of the multiplicative processes represented in an observable are relevant to a study. Attenuation by the ISM is a confounding variable for a study of a stellar populations based on a color magnitude diagram; the intrinsic spectral energy distribution of a star is a confounding variable in a study of the shape of the interstellar extinction curve. When using Bayesian inference to learn about the relevant processes, it is necessary to marginalize over the parameters describing the confounding processes. Because this marginalization can be time consuming if done using a numerical method such as Markov chain

Corresponding author: Kirill Tchernyshyov
ktcherny@gmail.com

Monte Carlo (MCMC), it is advantageous to marginalize over nuisance parameters analytically whenever this is possible. We have created a package, `name`[1], for analytically marginalizing over nuisance parameters that enter an observable multiplicatively and linearly (multiplicative linear nuisance parameters or MLNPs).

The problem this package is designed for is the analysis of absorption features in spectra. The process by which a spectrum containing absorption features is formed is the first example we gave in the paragraph above: a source emits light which is then attenuated by intervening matter. The unattenuated light is customarily called *the continuum*. To determine parameters describing the intervening matter, it is necessary to also determine parameters describing the continuum. While it is possible to pre-determine the continuum parameters using a portion of the spectrum that does not contain absorption features, we prefer to infer the absorption feature and continuum parameters simultaneously in order to explicitly include the uncertainty in continuum placement in estimates of the absorption feature parameters. This is also the approach used in the recently released absorption line analysis packages `Starfish`, `sick`, and `BayesVP` (Czekala et al. 2015; Casey 2016; Liang et al. 2018). In these packages, absorption feature and continuum parameters are simultaneously marginalized over using MCMC. The authors of both packages point out that marginalizing over large numbers of continuum parameters in this way can lead 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.

If the continuum is a linear function of the continuum parameters, the prior on the continuum parameters is (multivariate) normal or improper uniform, and the likelihood function of the observed spectrum given a model spectrum is multivariate normal, the necessary marginalization can be done analytically instead of numerically. The key is that when the absorption feature parameters are held fixed, the continuum basis elements, the transmittance spectrum, and, if necessary, the line spread function (LSF) can be combined into a single design matrix. The problem becomes equivalent to

[1] available *here*

the simple linear nuisance parameter model discussed in an astronomical context in Luger et al. (2017). The likelihood marginalized over the continuum parameters is available in e.g. Rasmussen & Williams (2006) for both the normal prior and the improper uniform prior cases.

Analytic marginalization over multiplicative linear nuisance parameters has been used for other problems in astronomy. In one common use case, the parameter that is marginalized over is the amplitude of a signal whose shape is a non-linear function of some parameters of interest. One example of this use case is Price-Whelan et al. (2017), where the parameters of interest are the period, eccentricity, phase, and argument of the radial velocity curve of a potential binary system and the nuisance parameters are the velocity amplitude of the curve and the barycenter velocity of the system. Another example can be found in Leistedt & Hogg (2017), where the parameter of interest is the redshift of a galaxy and the nuisance parameter is the galaxy's luminosity. This problem is not identical to the ones discussed in our paper and in Price-Whelan et al. (2017) because the luminosity multiplies both the data and the covariance matrix of the model residuals. The resulting marginal likelihood can not be expressed in terms of commonly used functions. The concept, however, is analogous and an accurate analytic approximation is available.

When developing this package, we were specifically considering properties of the absorption spectrum likelihood function and useful ways in which it can be factored. The mathematical contribution of this work is an expression for the gradient of the logarithm of the marginal likelihood with respect to the problem's interesting non-linear parameters. This gradient can be used for optimization or in gradient-aware sampling methods such as Hamiltonian Monte Carlo (Duane et al. 1987). The other contribution is the package itself, which uses the structure of the likelihood function to efficiently calculate the marginal likelihood and gradient. We state the problem and give expressions for the marginal likelihood and the gradient of its logarithm in Section 2. In Section 3, we describe features supported by the package and demonstrate its performance on representative test problems. We discuss strengths and weaknesses of our approach and package in Section 5 and conclude in Section 6.

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 $\mathbf{K}$:

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

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:

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

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 $\mathbf{L}\mathbf{B}\hat{\mathbf{c}}$ 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.

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

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

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

where

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

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

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

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

These 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_g is always proper because $\mathbf{C}_g$ is always positive definite. $\mathbf{C}_g$ is always positive definite because $\mathbf{\Lambda}^{-1}$ 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 $\mathbf{d}(\theta)$:

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

$$= \sum_{i=1}^{N} \left[\mathbf{J}^{i,i}\mathbf{A}_\mathbf{m} \;\; 0 \times \mathbf{A}_\mathbf{b}\right] \tag{18}$$

$$= \left[\mathbf{A}_\mathbf{m} \;\; \mathbf{0}\right], \tag{19}$$

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 exponential. 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. IMPLEMENTATION AND DEMONSTRATION

In this section, we describe some capabilities and limitations of the package (Section 3.1), show how the computation time of different calculations grows with dataset and basis size (Section 3.2), and compare the relative efficiencies of MCMC and analytic marginalization in test problems involving generated (Section 4.3) and real data (Section **??**).

3.1. *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 $\boldsymbol{\Lambda}$ 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.

3.2. *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 $\mathcal{O}(M^3)$ 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 dataset 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, 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 1, 2, and 3. 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 1). 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 1 and 3 should apply.

4. PRACTICAL TEST CASES

- Is marginalization of continuum parameters, analytic or not, actually useful? - We consider two metrics by which it could be useful/not useful: 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: a simpler one in which the continuum parametrization is known and more complicated one in which it is necessary to choose a continuum parametrization from a set of possibilites.

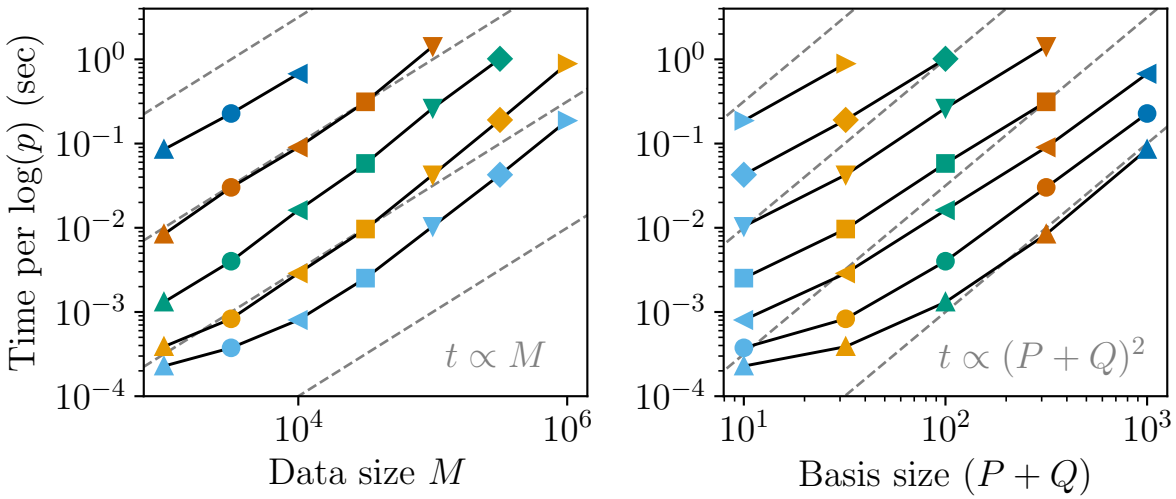

Figure 1. Computation time of the marginal log-likelihood (Equations 11 and 14) when the data covariance matrix $\mathbf{K}$ is diagonal and $\mathbf{L}$ is the identity mapping as a function of dataset size M (left panel) and basis size $P+Q$ (right panel). Values with the same marker shape were computed 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.

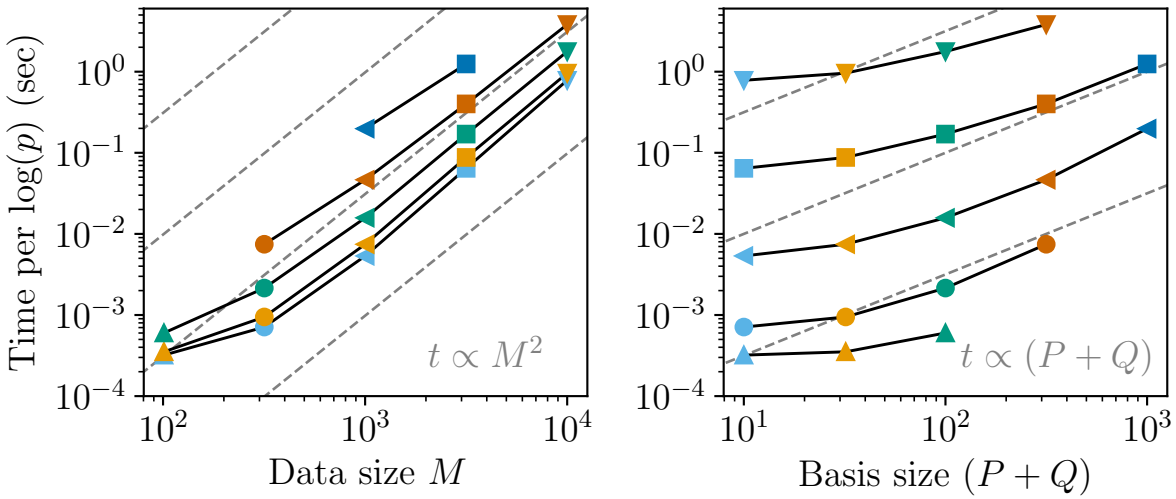

Figure 2. Computation time of the marginal log-likelihood when the data covariance matrix $\mathbf{K}$ is not diagonal and $\mathbf{L}$ is a dense matrix. See caption of Figure 1 for a description of figure elements.

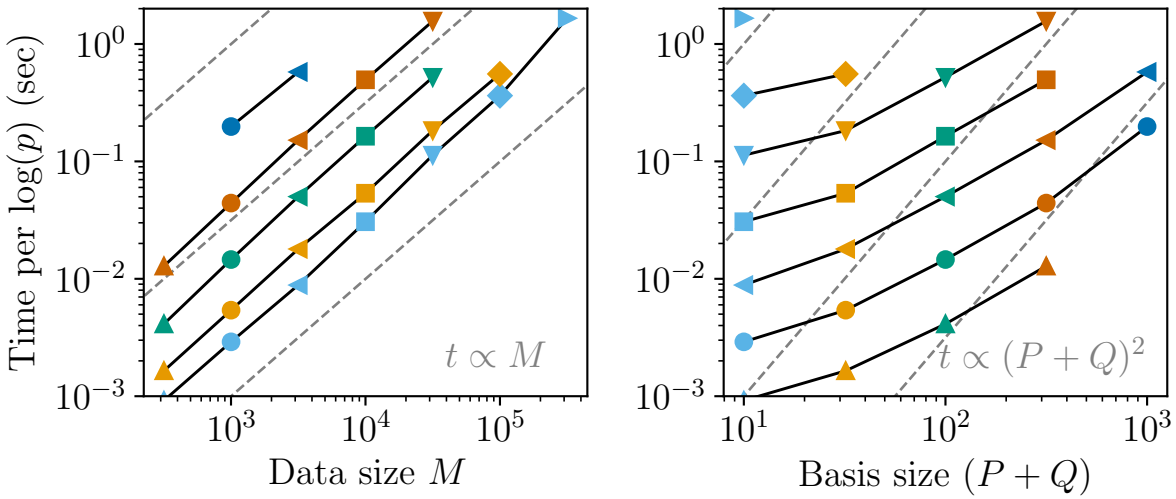

Figure 3. Computation time of the marginal log-likelihood when the data covariance matrix **K** is diagonal and **L** is a sparse, banded matrix. See caption of Figure 1 for a description of figure elements.

To demonstrate the usefulness of analytic marginalization in practical applications, we consider three test problems. The first problem is to determine the column density of an absorption line when continuum parameters are not known but the form of the continuum is known. The second problem is to determine a column density when the continuum parametrization is not known but the family of possible parametrizations is known. In these problems, we compare the accuracy and precision that can be attained using different fitting strategies. The third problem is to generate samples from a posterior probability distribution over absorption line parameters when the continuum parametrization is known. In this problem, we compare the computational efficiency of sampling with and without analytic marginalization over continuum parameters.

We are considering two types of use cases in these comparisons. The first two problems are relevant for spectroscopic surveys with too many spectra to analyze manually. The comparison is between different ways of automatically treating the continuum. The third problem is relevant for careful Bayesian analysis of a small number of spectra. In particular, we are considering a case where absorption lines are spread over different spectra or segments of a spectrum. We find that continuum

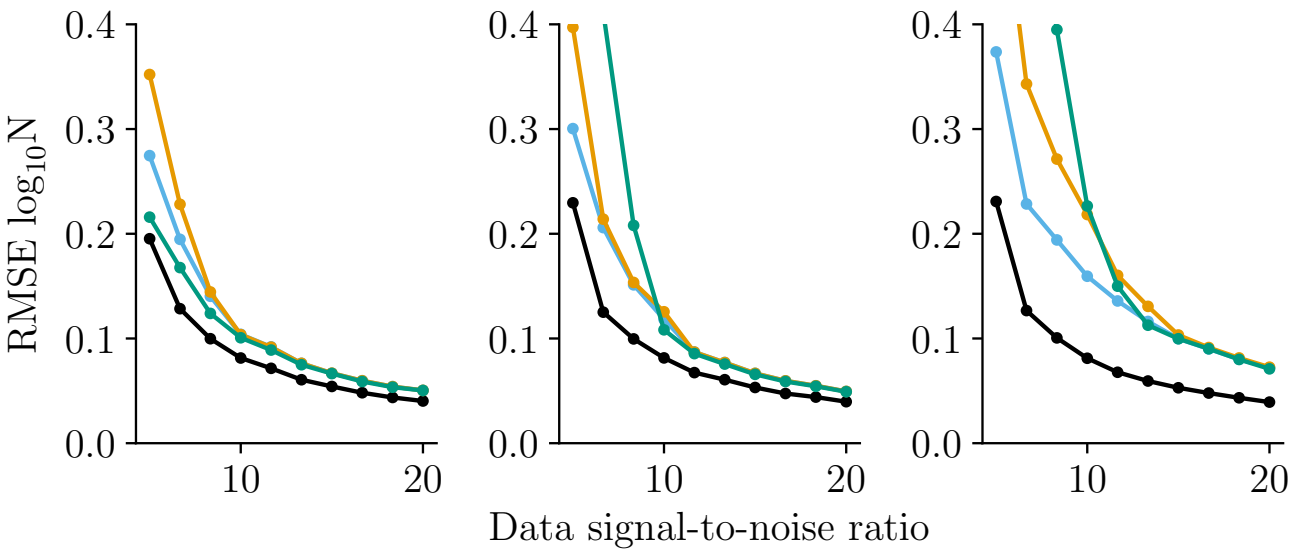

Figure 4. RMSE of different ways of estimating column densities assuming a known continuum parametrization.

marginalization can increase the accuracy and precision of automated spectral line analysis and that analytic continuum marginalization can, in some cases, help accelerate sampling.

4.1. *Marginalization over parameters*

— We consider the problem of measuring the column density of a single well-resolved, unsaturated absorption line on a continuum whose parametrization is known but whose parameters are not known. — This is the simplest problem that involves both absorption line and continuum parameters. — 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 summary statistic that we will use to indicate quality is the root mean square error (RMSE) of the base ten logarithm of the column density ($\log_{10}$ N). — We use the logarithm because physical constants (e.g. oscillator strengths) cancel in that RMSE calculation and because it is the logarithm of the column density that is usually reported.

— 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 accounting for 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 (but not translate that into a number), can perfectly estimate the best-fit absorption line profile by eye given this continuum, and using this profile can determine which part of the spectrum is not affected by the line.

— The continuum parametrizations we use are polynomials of order 0, 1, and 2. — We consider SNRs between 5 and 25. — To compute each RMSE, we generate 1000 spectra. — All spectra contain the same absorption line but have different continuum parameters and noise realizations.

— The RMSEs obtained using these different methods are shown in Figure 4. — 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 human-like analysis. — The advantage of marginalization over the other methods becomes greater as the continuum parametrization becomes more complex. — *Marginalization is always the most robust (in a statistical sense) method.*

4.2. *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. — We consider a different range in SNR: 10 to 100.

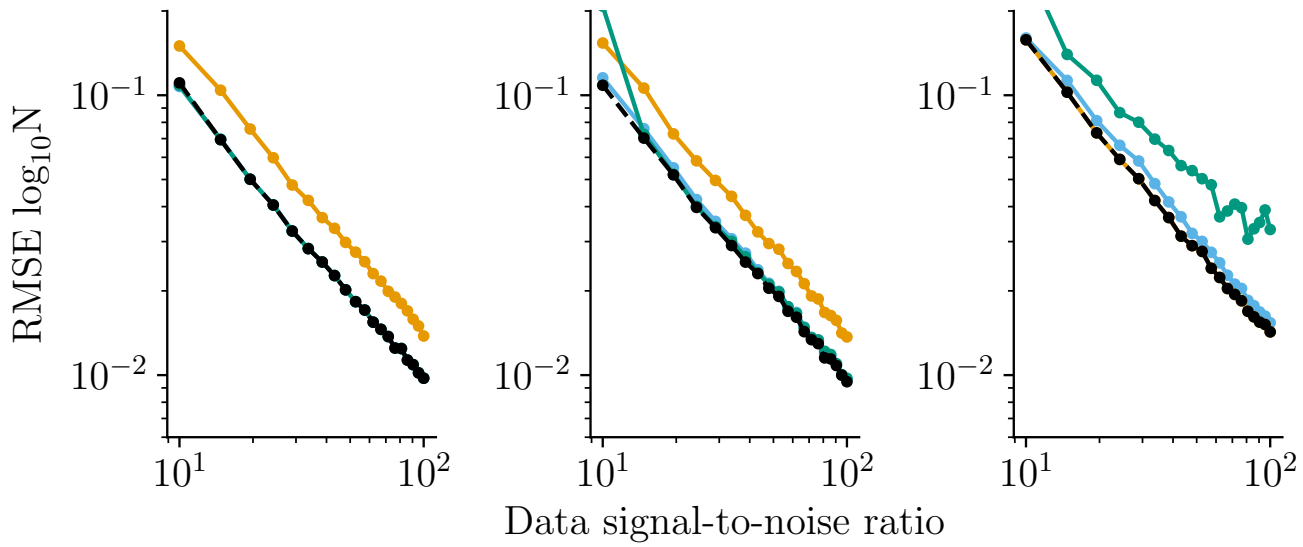

Figure 5. Different ways of dealing with an unknown continuum parametrization

— 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 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 parametrization and in other (4) essentially through model averaging. — 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 5. — Methods (1), (2), and (4) are "robust" in the sense that their RMSE depends on SNR in a consistent way. — Method (3), parametrization selection, is not robust in this sense. — At an SNR of 10 for spectra with 1st order continua and

at all of the considered SNRs for spectra with 2nd order continua, parametrization selection fails. — In all of the other cases, parametrization selection and marginalization are indistinguishable. — Given these two facts, marginalization is clearly the more effective of the two multi-parametrization methods.

— For all three input orders, the parametrization-marginalized solution is only slightly worse than the reference, known-parametrization solution. — Its RMSE is, on average, TKTK, TKTK, and TKTK higher than the reference solution for the 0th, 1st, and 2nd order spectra. — For spectra generated with the simpler continuum parametrizations, the RMSE of the conservative method is higher than the RMSE of parametrization-marginalized solution by TKTK and TKTK. — When the order is 2, the conservative solution is the same as the reference solution. — All of these ratios are approximately constant across this entire SNR range.

— A different summary statistic is the ratio of the SNRs required by the different methods in order to achieve the same RMSE. — For 0th, 1st, and 2nd order true parametrizations, the conservative required SNR divided by the parametrization marginalized SNR is TKTK, TKTK, and TKTK. — Assuming the SNR of these spectra is proportional to the square root of integration time, parametrization marignalization requires TKTK and TKTK times less time for the 0th and 1st order cases and TKTK more time for the 2nd order case.

4.3. *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 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

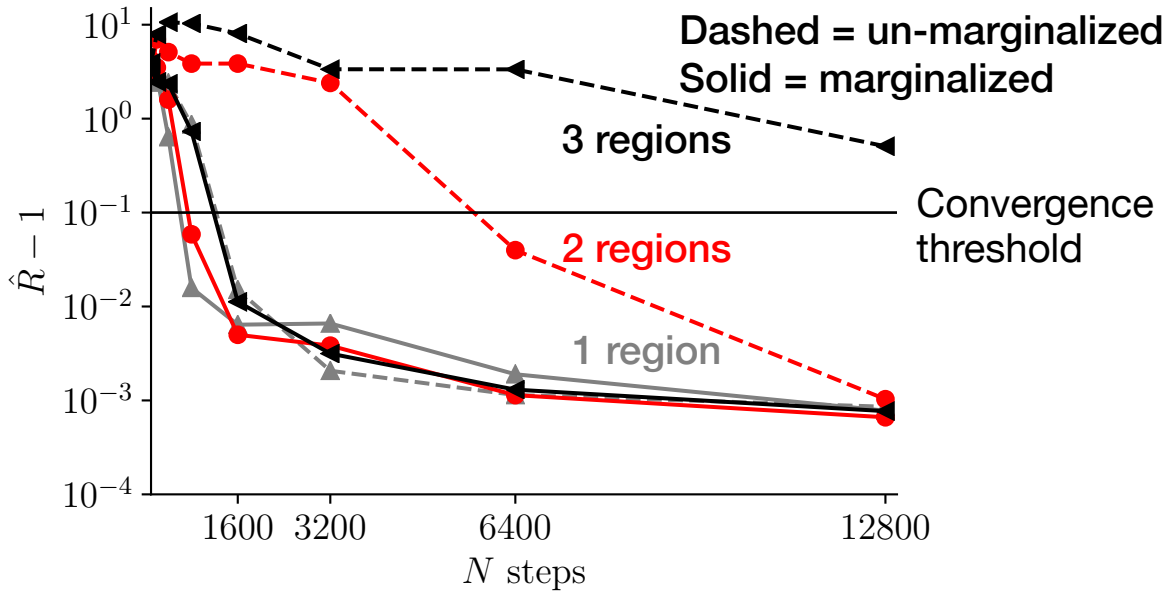

Figure 6. Rubin-Gelman statistic

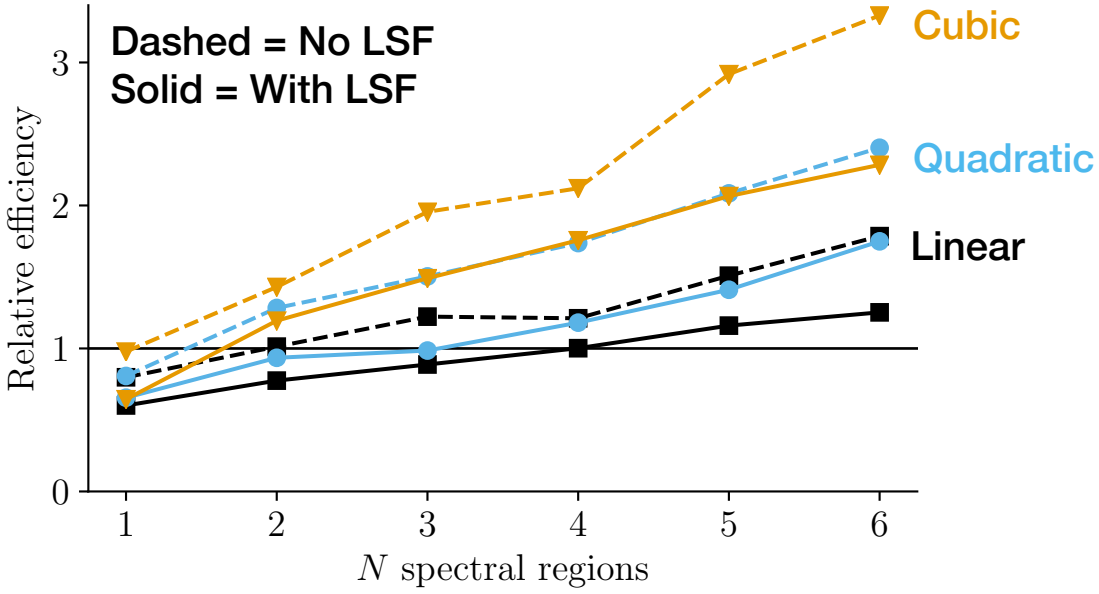

Figure 7. Independent samples per second

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.

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. 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 6. 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 7. 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 3.2).

5. DISCUSSION

5.1. *Assumptions and consequences*

- Assumptions made are Gaussianity, linearity, un-constrained, known covariance/uncertainty - Trivially does not apply to any absorption analysis problem; no background source produces negative flux. - Less tedious case is: observations in the low count rate, e.g. Poisson regime. - Poisson data break Gaussianity, not being constrained, and (if you're using sqrt(intensity) rather than a pre-assumed uncertainty) the covariance assumption. - First two mean you're doing an approximation; last one actually rules out using this particular analytic marginalization formula (though see work by Leistedt and Hogg, where just the covariance/uncertainty assumption is relaxed) - Importance of Gaussian vs. Poisson is just the usual "how good of an approximation is one distribution to the other" question - Basically, going to be in trouble with low-count X-ray or UV spectra; most optical and longer spectra are going to be fine.

- Continuum model envisioned here is linear because it's just an effective estimate that ignores (generally non-linear) physics - this will be fine in some applications: the simple continua of quasars, some rapidly rotating O stars; any time that the (pseudo-)continuum changes on wavelength scales that are substantially longer than the wavelength scale of the lines - this is possibly longer than you think; P Cygni profiles from stellar winds are funky, but can be described reasonably well by splines - Need models that are informed by physics for other cases: optical or NIR spectra of cooler stars (e.g. Zasowski) - MLNPs may still be useful here given that models of these spectra are not perfect; small corrections may be necessary and particularly important for weak lines!

- An assumption that is not necessary for analytic marginalization to be possible but is necessary for it to be useful is that the span of the absorption model includes the actually observed absorption. - In simpler terms, if you model a spectrum that contains two distinct lines with a single line, continuum marginalization will not save you. : Where do I put this comment? It's too short to be a separate paragraph and doesn't really fit into the other two paragraphs. Maybe the statement is: also need correct-enough model of the absorption itself? Hm.

5.2. *Towards automated absorption line analysis*

- The test cases in Section 4 showed that continuum marginalization in general and analytic marginalization in particular has some advatanges over other approaches to accounting for continua. - Here, we explore how these advantages could be applied in two use cases: automated but well-motivated analysis of lots of simple spectra and thorough, reproducible analysis of single complicated spectra. - In both cases, interested in making it possible to do *blind* and reproducible analysis. - When testing hypotheses, good to do as much of the analysis as possible before looking at results to discourage tinkering with analysis to produce more

desirable results. - Reproducibility is also a good idea and is not directly possible if people are analyzing spectra manually and actively participating in continuum placement.

- to extract ISM absorption information from e.g. the SDSS spectroscopic surveys, two things need to be automated: finding absorption lines and accounting for the continuum. - this has always been done with fitting rather than marginalization, and selection of continuum parametrizations has not been done. - as was shown in one of the test cases, choosing a conservative parametrization leaves SNR on the table. - analytic marginalization means that one can do optimization for line parameters and marginalization and even parametrization selection for continuum parameters. - useful for samples of objects with heterogenous spectra, e.g. GALAH which has some hot main sequence stars and some cool giants; automatically use a simple model for the hot stars and more complicated model for the cool stars. - automation of finding absorption lines is more complicated. - past cases of line analyses in surveys have used prior information, e.g. the existence of a galaxy along the line of sight (citation). - to make blind, statistically sound ISM studies from large spectroscopic surveys possible, still need a way of locating lines; this is tricky, since absorption is non-linear.

- Needs of careful analysis of individual spectra are the same: automate location of absorption lines and continuum placement. - Since you have only one spectrum, can use more time consuming methods; do inference rather than optimization. - Way of probabilistically placing lines already exists! "Trans-dimensional inference," where the dimensionality of the problem, e.g. the number of absorption lines, is a parameter to be inferred. - e.g. "Reversible Jump MCMC," which I used in CITET DEPLETION PAPER. - Particularly important for when instrument resolution is not sufficient to resolve finest lines present; unresolved structure can hide column density. - Having analytic continuum marginalization & continuum parametrization selection here means that probabilistic analysis of individual spectra/small samples really can be automated and blinded.

6. CONCLUSION

have a package that computes marginal likelihoods and their gradients idea of marginalizing over linear parameters is not new, but to best of our knowledge had not been applied to problem of absorption line continua. there are some problem-specific features, e.g. LSFs, which can be handled more efficiently with a purpose-built package.

have carried out tests on whether continuum marginalization is any better than several simpler approaches. have shown that in terms of accuracy of recovered parameters, marginalization is best way of doing "simple," "local" continua. true on two levels: if you have a good guess as to the correct parametrization, marginalizing over the parameters will, at low SNR, get you more accurate parameter recovery. if you only have a

good guess as to the set of plausible parametrizations, marginalizing over parametrizations and parameters within each parametrization is almost as good as knowing the correct parametrization. marginalizing over parametrizations will, on average, yield more accurate parameter values than any of the other strategies we tested. (do I need to test picking a parametrization using a likelihood test?) this extra accuracy is free! it's like having a higher SNR spectrum and analyzing it in the simpler, less-good way. this extra accuracy does not require much extra compute and (when line placement is known) can be used automatically, making it great for analysis of data from massive spectroscopic surveys.

have also considered an "opposite" case, a single spectrum with many different lines that you want to analyze very carefully with MCMC. analytic continuum marginalization can dramatically speed up convergence of MCMC and shorten chain autocorrelation times (since there are fewer parameters to deal with numerically). when there are many continuum parameters to deal with (as is the case when you have a bunch of splines, e.g.), shortening of autocorrelation times can mean more independent samples per second—greater efficiency.

list of results:

- likelihood and gradient
- package implementing the above
- demonstration that continuum marginalization is practically useful in terms of achieving greater accuracy and, in some cases, efficiency

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

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

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

Leistedt, B., & Hogg, D. W. 2017, ApJ, 838, 5

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