

**An example of statistics in high-resolution X-ray spectroscopy -
velocity profiles in the O-star supergiant ζ Ori**

**Andy Pollock
European Space Agency
XMM-Newton RGS Calibration Scientist**

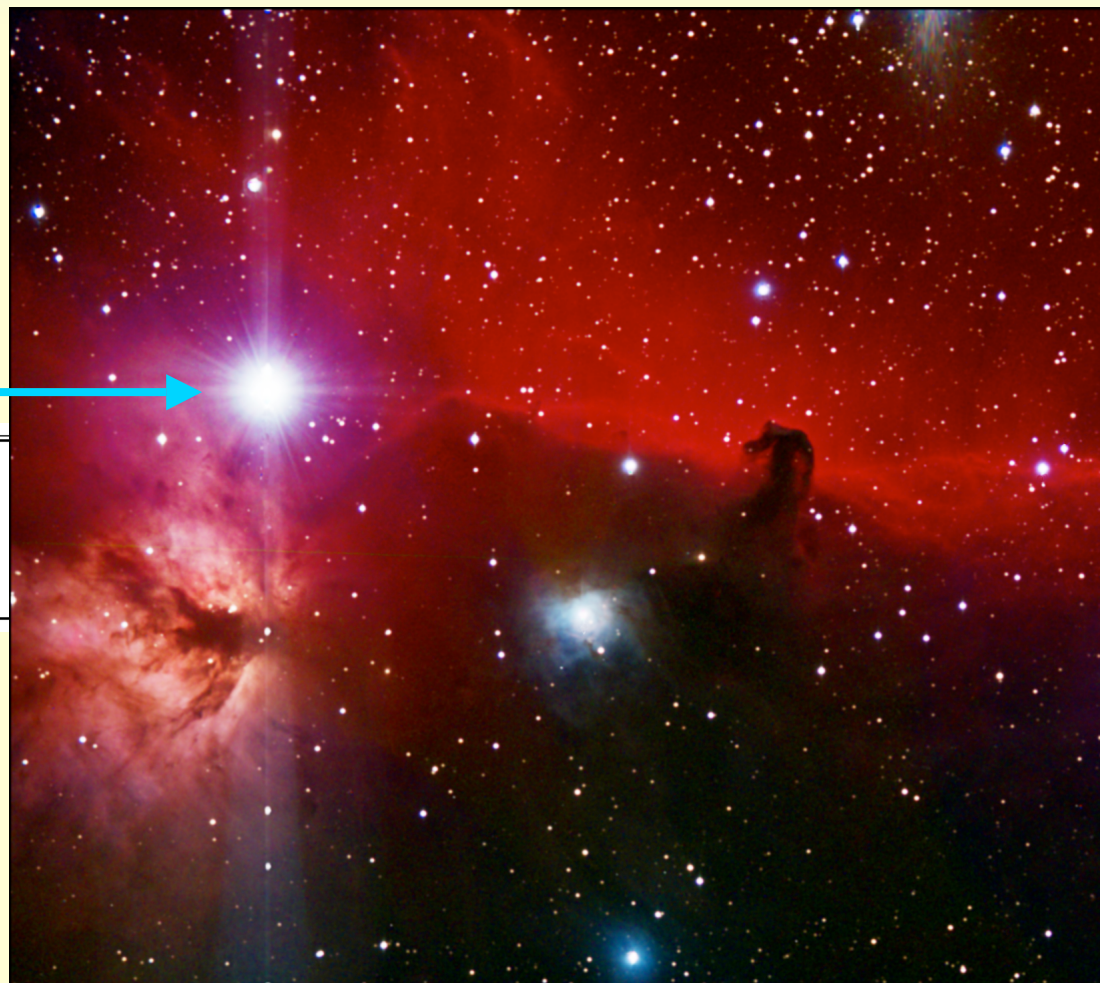
MSSL School in High-Resolution Spectroscopy

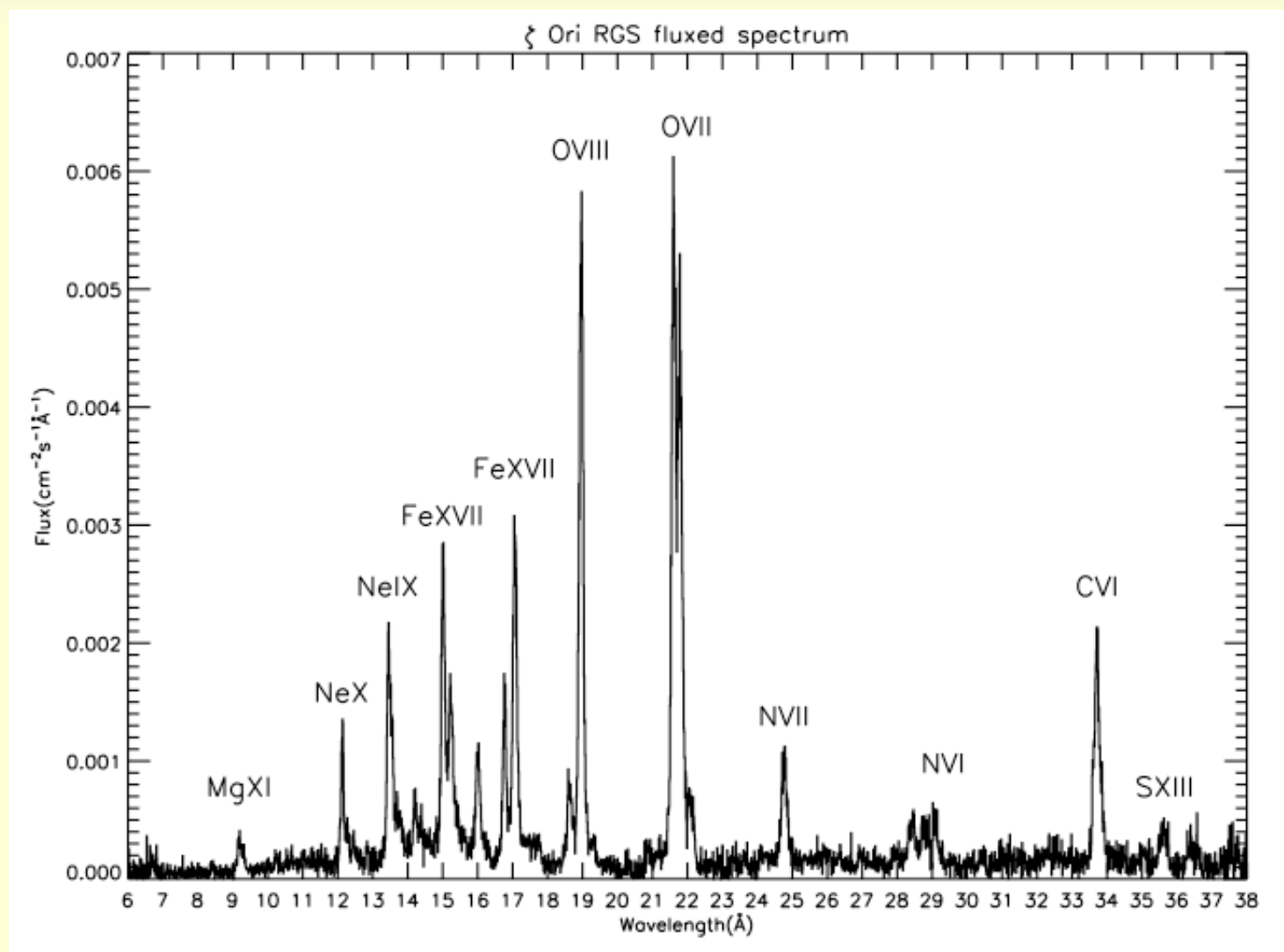
2009 March 17

Make every photon count.
Account for every photon.

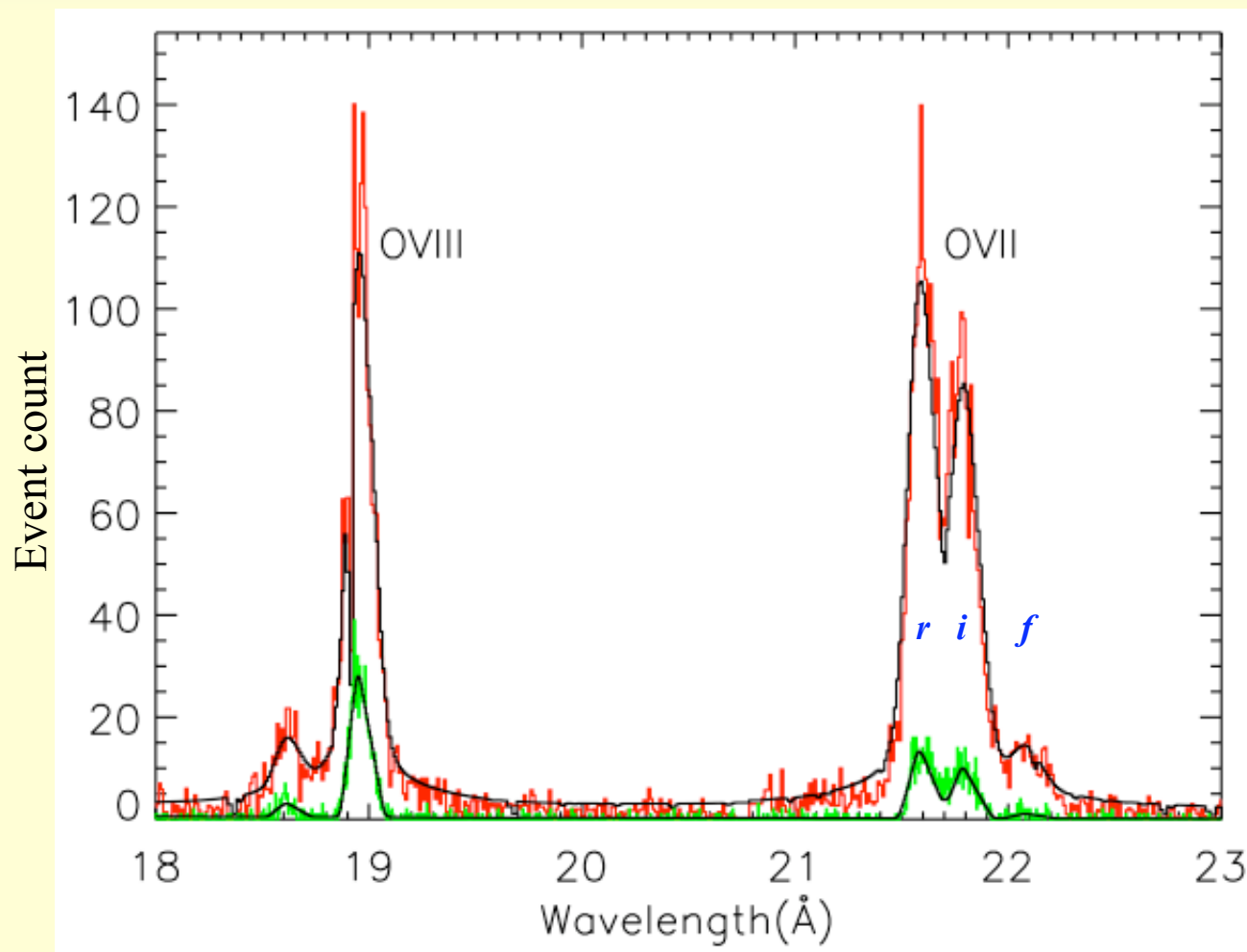
HD37742 O9.7Ib ζ Orionis

Visual magnitude	V	2.03
Effective temperature	T_{eff}	30 900 K
Radius	R_*	$31 R_{\odot}$
Terminal velocity	v_{∞}	2100 km s^{-1}
Mass-loss rate	$\dot{M}$	$1.4 \times 10^{-6} M_{\odot} \text{ yr}^{-1}$
Distance	d	473 pc
Interstellar column density	N_{H}	$2.5 \times 10^{20} \text{ cm}^{-2}$





Part of the high-resolution X-ray spectrum of ζ Ori



XMM RGS
Chandra MEG

Instrument	Order	Counts
RGS1	-1	15 136
RGS2	-1	14 133
RGS1	-2	2771
RGS2	-2	2507
MEG	-1	5556
MEG	+1	3568
HEG	-1	1129
HEG	+1	839

data $\Leftrightarrow$ models

$$\{n_i\}_{i=1,N} \Leftrightarrow \{\mu_i\}_{i=1,N}$$

≥ 0 individual events $\Leftrightarrow$ continuously distributed

detector coordinates $\Leftrightarrow$ physical parameters

never change $\Leftrightarrow$ change limited only by physics

have no errors $\Leftrightarrow$ subject to fluctuations

most precious resource $\Leftrightarrow$ predictions possible

kept forever in archives $\Leftrightarrow$ kept forever in journals and textbooks



There are two sorts of statistic(s)

- Gaussian statistics $\Leftrightarrow \chi^2$ -statistic
- Poisson statistics $\Leftrightarrow$ C-statistic

Likelihood of data on models



$$L = \prod_{i=1}^N P(n_i | \mu_i)$$

Gaussian

$$L = \prod_{i=1}^N \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(n_i - \mu_i)^2}{2\sigma_i^2}\right) dn_i$$

$$\ln L = -\frac{1}{2} \sum_{i=1}^N \frac{(n_i - \mu_i)^2}{\sigma_i^2} - \sum_{i=1}^N \ln \sigma_i + \kappa(\ln dn_i)$$

$$-2 \ln L = \chi^2$$

Poisson

$$L = \prod_{i=1}^N \frac{e^{-\mu_i} \mu_i^{n_i}}{n_i!}$$

$$\ln L = \sum_{i=1}^N n_i \ln \mu_i - \mu_i - \kappa(\ln n_i!)$$

$$-2 \ln L = C$$

Cash 1979, ApJ, 228, 939

The model

$$\mu(\underline{\theta}, \underline{\beta}, \underline{\Delta}, \underline{D}) = S(\underline{\theta}(\underline{\Omega})) \otimes R(\underline{\Omega} < \underline{\Delta} > \underline{D}) + B(\underline{\beta}(\underline{D}))$$

$\underline{D}$ = set of detector coordinates $\{X, Y, t, PI, \dots\}$

S = source of interest

$\underline{\theta}$ = **set of source parameters**

R = instrumental response

$\underline{\Omega}$ = set of physical coordinates $\{\alpha, \delta, \tau, v, \dots\}$

$\underline{\Delta}$ = set of instrumental calibration parameters

B = background

$\underline{\beta}$ = set of background parameters

$$\Rightarrow \ln L(\underline{\theta}, \underline{\beta}, \underline{\Delta}) \Rightarrow \ln L(\underline{\theta})$$

Uses of the log-likelihood, $\ln L(\theta)$

- $\ln L$ is what you need to assess all and any data models
 - locate the maximum-likelihood model when $\underline{\theta} = \underline{\theta}^*$
 - minimum χ^2 is a maximum-likelihood Gaussian statistic
 - minimum C is a maximum-likelihood Poisson statistic
 - compute a goodness-of-fit statistic
 - reduced chi-squared $\chi^2/\nu \sim 1$ ideally
 - reduced C $C/\nu \sim 1$ ideally
 - ν = number of degrees of freedom
 - estimate model parameters and uncertainties
 - $\ln L(\underline{\theta})$
 - $\underline{\theta}^* = \{p_1, p_2, p_3, p_4, \dots, p_M\}$
 - investigate the whole multi-dimensional surface $\ln L(\underline{\theta})$
 - compare two or more models
- calibrating $\ln L$, $2\Delta \ln L \Leftrightarrow \sigma \sqrt{2\Delta \ln L}$
 - $2\Delta \ln L < 1$. is not interesting
 - $2\Delta \ln L > 10$. is worth thinking about (e.g. 2XMM DET_ML ≥ 8 .)
 - $2\Delta \ln L > 100$. Hmmm...

Goodness-of-fit

- Gaussian model and data are consistent if $\chi^2/\nu \sim 1$
 - ν = “number of degrees of freedom”
= number of bins – number of free model parameters
= $N - M$
 - $cf < (x - \mu)^2 / \sigma^2 > = 1$
 - same as comparison with best-possible $\nu=0$ model, $\underline{\mu} = \underline{x}$,
 - $\chi^2 = 2(\ln L(\underline{\mu} = \underline{x}) - \ln L(\underline{\theta}))$
- Poisson model and data are consistent if $C/\nu \sim 1$
 - comparison with best-possible $\nu=0$ model, $\underline{\mu} = \underline{n}$
 - $2\sum(n_i \ln n_i - n_i) - 2\sum(n_i \ln \mu_i - \mu_i) = 2\sum n_i \ln(n_i / \mu_i) - (n_i - \mu_i)$
 - XSPEC definition
 - What happens when many $\mu_i \ll 1$ & $n_i = 0$?

Estimate model parameters and their uncertainties

- Parameter error estimates, $d\underline{\theta}$, around maximum-likelihood solution, $\underline{\theta}^*$
 - $2\ln L(\underline{\theta}^* + d\underline{\theta}) = 2\ln L(\underline{\theta}^*) + 1$. for 1σ (other choices than 1. sometimes made)

Practical considerations

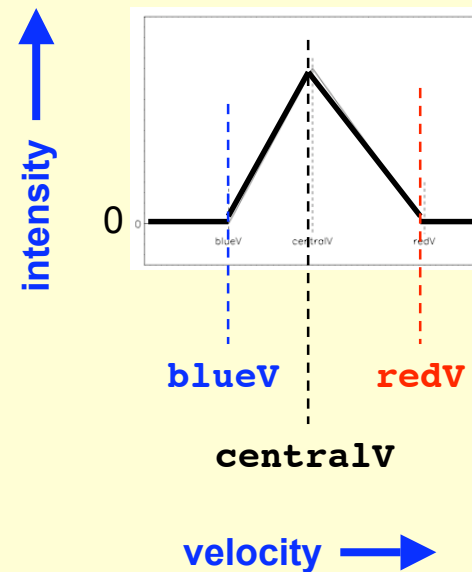
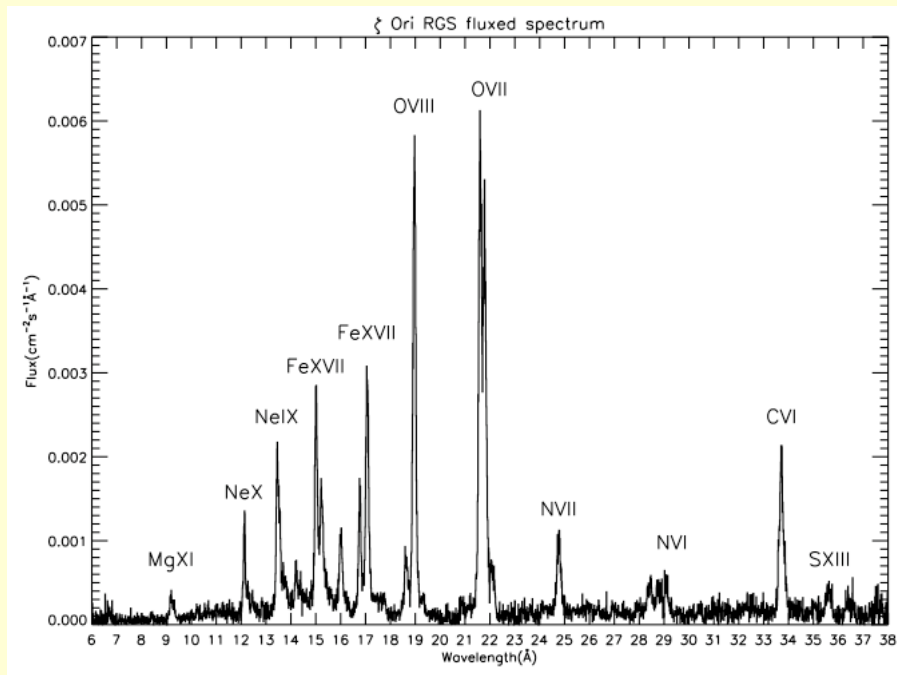
- S/ν is rarely ~ 1
 - $S = \chi^2 |C|$
 - $\ln L(\underline{\theta}, \underline{\beta}, \underline{\Delta})$
 - $\underline{\theta}$ = set of source spectrum parameters
 - physics might need improvement
 - $\underline{\beta}$ = set of background parameters
 - background models can be difficult
 - $\underline{\Delta}$ = set of instrumental calibration parameters
 - 5 or 10% accuracy is a common calibration goal
- solution often dominated by systematic errors
 - XSPEC's SYS_ERR is the wrong way to do it
 - no-one knows the right way (although let's listen to those Gaia people...)
- formal probabilities are not to be taken too seriously
 - $S/\nu > 2$ is bad
 - $S/\nu \sim 1$ is good
 - $S/\nu \sim 0$ is also bad
- find out where the model isn't working
 - pay attention to every bin

To rebin or not to rebin a spectrum ?

- Pros
 - Gaussian $\equiv$ Poisson for $n \gg 0$
 - dangers of oversampling
 - saves time
 - everybody does it
 - “improves the statistics”
 - grppha and other tools exist
 - on log-log plots $\ln 0 = -\infty$
- Cons
 - rebinning throws away information
 - 0 is a perfectly good measurement
 - images are never rebinned
 - Poisson statistics robust for $n \geq 0$
 - $\mu_1 + \mu_2$ is also a Poisson variable
 - oversampling harmless
 - adding bins does not “improve the statistics”

Leave spectra alone. Don't rebin for $\ln L(\theta)$. Use Poisson statistics.

- instead of Gaussian lines use a **TriLine** model
- $\mu = 67 \times \text{TriLine} + \text{continuum}$



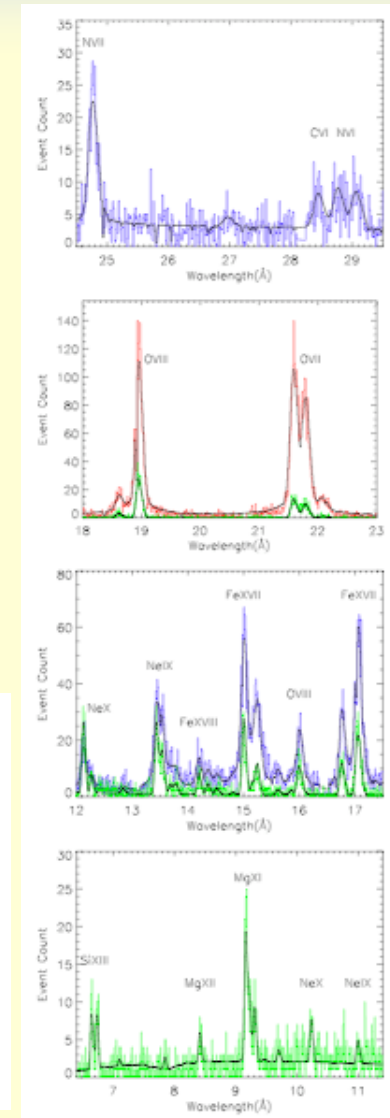
- use the model that best suits your purpose
- ...and if there isn't one, make one of your own
 - scripts
 - XSPEC Tcd
 - ISIS Slang
- use **all** the data available

1 σ errors for $\Delta C=1$ with XSPEC> error 1. ...

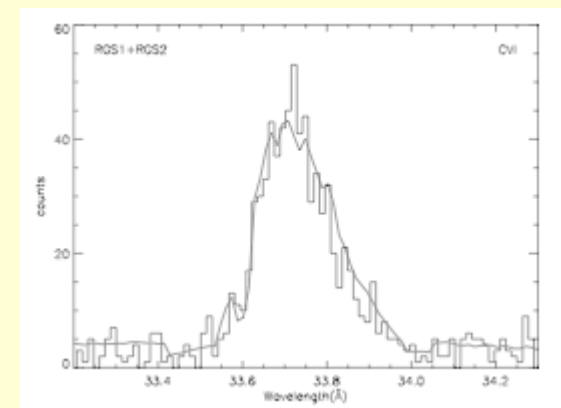
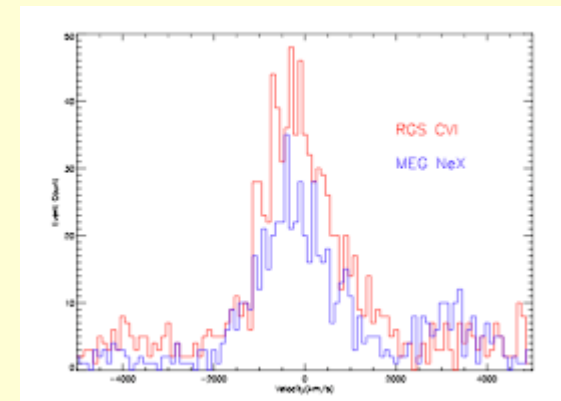
Table 3. ζ Orionis best-fit emission-line velocity parameters and underlying continuum for the simultaneous multiple common-profile TriLine XSPEC model fit to the *XMM-Newton* RGS 1st and 2nd order spectra of 2002-09-15 and *Chandra* MEG and HEG ± 1 order spectra of 2000-04-08.

TriLine velocities		
blueV	-1642 ± 22	km s $^{-1}$
centralV	-302 ± 29	km s $^{-1}$
redV	$+1646 \pm 26$	km s $^{-1}$
bremsstrahlung continuum		
N_H	2.5×10^{20}	cm $^{-2}$
kT	0.494 ± 0.007	keV
normalisation	$5.66 \pm 0.14 \times 10^{-3}$	
C-statistic = 22 862.4 using 26281 PHA bins		

	blueV	centralV	redV	C-stat	constraint
TriLine	-1642 ± 22	-302 ± 29	$+1646 \pm 26$	22 864.4	Best fit of Table 3
TriLine	-1641 ± 15	-303 ± 26	$+1641 \pm 15$	22 865.9	$ \text{redV} = \text{blueV} $
TriLine	-1774 ± 21	0	$+1492 \pm 21$	22 973.1	centralV = 0
TriLine	-1684 ± 15	0	$+1684 \pm 15$	23 053.2	centralV = 0, $ \text{redV} = \text{blueV} $
	blueW	centralV	redW	C-stat	constraint
SkewLine	650 ± 25	-318 ± 29	991 ± 27	22 816.7	
SkewLine	811 ± 9	-126 ± 10	811 ± 9	22 879.4	redW = blueW
SkewLine	880 ± 12	0	738 ± 12	22 947.6	centralV = 0
SkewLine	841 ± 10	0	841 ± 10	23 038.4	centralV = 0, redW = blueW



ion	blueV (km s ⁻¹)	centralV (km s ⁻¹)	redV (km s ⁻¹)	C-stat	ΔC
All ions	-1642 ± 22	-302 ± 29	+1646 ± 26	22862.4	
C VI	-1719 ± 135	-270 ± 137	+1403 ± 130	22855.6	6.8
N VII	-1705 ± 55	-111 ± 240	+1065 ± 324	22859.8	2.6
O VII	-1349 ± 72	-509 ± 83	+1835 ± 76	22844.9	17.5
O VIII	-1656 ± 43	-258 ± 59	+1482 ± 46	22838.7	23.7
Fe XVII	-1647 ± 42	-421 ± 62	+1845 ± 59	22848.9	13.5
Ne IX	-1653 ± 77	-211 ± 79	+1508 ± 102	22859.9	2.5
Ne X	-1730 ± 106	-327 ± 126	+1735 ± 103	22861.0	1.4
Mg XI	-1536 ± 153	+33 ± 121	+1487 ± 144	22850.1	12.3
Mg XII	-721 ± 365	-662 ± 227	+1514 ± 471	22855.1	7.3
Si XIII	-1449 ± 124	+48 ± 240	+1421 ± 188	22856.4	6.0



Make every photon count.
Account for every photon.