Science from Gaia: how to deal with a complex billion-source catalogue and data archive

Anthony Brown

Sterrewacht Leiden, Leiden University brown@strw.leidenuniv.nl

- Gaia mission overview
- Complexities of the Gaia catalogue
- Modelling the Milky Way
- Extracting the maximum science from the data archive



ESA Cornerstone mission within Horizon 2000+ programme
Create large and highly accurate stereoscopic map of the Galaxy
Global astrometry concept successfully demonstrated by Hipparcos





Launch in 2013 with Soyuz-Fregat from Kourou
Orbit: vicinity of L2
Mission duration 5 (+1) years

Image credit: Lund Observatory







Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

SAFRAN

La Palma 30.05.2011 - p.5/47



Survey capabilities

- Three simultaneous observing modes
- Complete to G = 20 (V = 20-22)
- Observing programme: autonomous on-board detection and unbiased
- Quasi-regular time-sampling over 5 years (~ 80 observations)
- Angular resolution comparable to HST

Number of objects

- 1 billion stars to G = 20
- $10^6 10^7$ galaxies
- 500 000 quasars
- ◆ 3 × 10⁵ solar system bodies
- tens of thousands of exoplanets



La Palma 30.05.2011 - p.7/47

Astrometry

- $\sigma \sim 5-14 \ \mu \text{as } V < 12$, 10–25 μas V = 15, 100–300 $\mu \text{as } V = 20$
- $25\,000 \,\star/deg^2$; max $\sim 10^6 \,\star/deg^2$
- Census extra-solar planets to 200 pc
- ◆ 3 × 10⁵ minor bodies of the solar system, 100 masses

• PPN γ to $\sim 2 \times 10^{-6}$



Photometry

- Two channels: 330–680 nm (BP), 640–1000 nm (RP)
- Low resolution (~ 3–30 nm/pixel) prism spectra
- Allows derivation of A_V , T_{eff} , log g, [M/H], and [α /H] for brighter stars



Gaia variable star survey

- ~ 70 epoch survey over 5 years
- mmag accuracy per single observation

Quantitative impact

- 20×10^6 classical variables
- 1–5 million eclipsing binaries
- ~ 5000 Cepheids, 70 000 RR Lyr
 - RR Lyr visible out to ~ 75 kpc



Eyer & Mignard 2005

Spectroscopy

- Radial velocities
- Rotational velocities
- Atmospheric parameters
- Abundances
- Interstellar reddening

Diagnostics from spectroscopy

- Binarity/multiplicity, variability
- $\sim 10^6$ spectroscopic binaries
- $\sim 10^5$ eclipsing binaries ($\sim 25\%$ $SB2 \rightarrow masses$)
- Long period classical Cepheids $\sigma_{\nu_{\rm r}} < 7 \text{ km/s} \rightarrow 20\text{--}30 \text{ kpc}$

$$V \le 17 \quad \sim 150 \times 10^{6} \text{ stars}$$
$$V \le 13 \quad \sim 5 \times 10^{6}$$
$$V \le 13 \quad \sim 5 \times 10^{6}$$
$$V \le 12 \quad \sim 2 \times 10^{6}$$

$$V \le 12 \quad \sim 2 \times 10^6$$
$$V \le 13 \quad \sim 5 \times 10^6$$

 $V \leq$



La Palma 30.05.2011 - p.11/47

Other

- Accurate stellar classification for all classes and types
- Recalibration of the distance scale
- 10 000 stellar masses $\sigma < 1\%$
- 5×10^5 QSOs + z + photometry, ICRF in the visible

Potential Triggers for Gaia Alerts



Lukasz Wyrzykowski, IoA Cambridge UK

Transients with the Gaia Mission

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.13/47

Potential Triggers for Gaia Alerts



Workshop from June 29th to 1st July 2011

http://www.ast.cam.ac.uk/ioa/wikis/gsawgwiki/index.php/Workshop2011:main



Lukasz Wyrzykowski, IoA Cambridge UK

Transients with the Gaia Mission

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.14/47

Survey strategy



Survey strategy



 t_0 , $t_0 + 106$ minutes, $t_0 + 6$ hrs, $t_0 + 6$ hrs + 106 minutes, repeated 10–30 days later

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.16/47

Survey strategy



 t_0 , $t_0 + 106$ minutes, $t_0 + 6$ hrs, $t_0 + 6$ hrs + 106 minutes, repeated 10–30 days later

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.17/47

Hipparcos Catalogue statistics (1)



Image credit: ESA

Median number of observations (Ecliptic coordinates)

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.18/47

Hipparcos Catalogue statistics (1)



Image credit: ESA

Median duration between first and last observation (Ecliptic coordinates)

scanning law, mission duration, observation interruptions

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.19/47

Hipparcos Catalogue statistics (1)

Image credit: ESA

Median standard error on ϖ (Equatorial coordinates)

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.20/47

Correlated errors

Covariance matrix for measured quantity v:

$$\begin{aligned} \mathbf{C}_{\mathbf{v}} &= \mathrm{E}\left[(\mathbf{v} - \langle \mathbf{v} \rangle)(\mathbf{v} - \langle \mathbf{v} \rangle)^{T}\right] = \mathrm{E}\left[\Delta \mathbf{v} \Delta \mathbf{v}^{T}\right] \\ c_{ij} &= \rho_{j}^{i} \sigma_{i} \sigma_{j} \end{aligned}$$

Confidence region around $\langle \mathbf{v} \rangle$:

$$\Delta \mathbf{v}^T \mathbf{C}_{\mathbf{v}}^{-1} \Delta \mathbf{v} = z$$

Transformation:

$$\Delta \mathbf{w} = \mathbf{M} \Delta \mathbf{v} \to \mathbf{C}_{\mathbf{w}} = \mathbf{M} \mathbf{C}_{\mathbf{v}} \mathbf{M}^{T}$$

Example:

$$\begin{pmatrix} \sigma_{\lambda*}^2 & \rho_{\varpi}^{\lambda*}\sigma_{\lambda*}\sigma_{\varpi} \\ \rho_{\varpi}^{\lambda*}\sigma_{\lambda*}\sigma_{\varpi} & \sigma_{\varpi}^2 \end{pmatrix}$$

Hipparcos Catalogue statistics (2)

Image credit: ESA

Median correlation between $\lambda *$ and ϖ (Ecliptic coordinates)

• Systematics from asymmetric distribution of observations wrt position of Sun

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.22/47

Hipparcos Catalogue statistics (2)

Image credit: ESA

Median correlation between $\mu_{\lambda*}$ and μ_{β} (Ecliptic coordinates)

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

Hipparcos Catalogue statistics (2)

Image credit: ESA

Median correlation between $\mu_{\alpha*}$ and μ_{δ} (Equatorial coordinates)

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.24/47

Effect of coordinate transformation

Correlation between $\alpha *$ and δ due to coordinate transformation (Ecliptic to Equatorial)

- Assume $\lambda *$ and β uncorrelated
- $\sigma_{\lambda*}/\sigma_{\beta}$ varies with β

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

Star to star correlations

How to deal with these in Gaia case: presentation by Berry Holl

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.26/47

The Milky Way from Gaia

Turn this

1896

101 119 305 - 101 120 465

M301	- 94400					_						_		
Summer	Convipto		International Contention of Co	Pasilion	aproxits 199001-201	Par-	Proper	Making:	- 14	nini i	in sea		Lánmán Cardáire (k)	Rain.
		-72-	÷.	1. 21		÷	- C	40 I	2.0	÷	1	£.	221277277 <u>2</u>	÷
14.101	-	- 84 14 19 1	441	ter and the T					-			1.0	of a second second second	
1112	111.20	:2025	22	3125.014	: (b) (c) (c) (c)	110	23	22	8.0	14	10	100		1.11
-18.201	1411-0-36	1124-045	14	241.626.621.6	.18.810 201 IB		- 10	16.14	10.00	1.5	100	12	- k-11 + 1 + R - N - 1 - R - 11 - 5 - 51	5 85
104.125	1411-0-34	18113182	141	• 261 size 121 si	100-001706-01		1.01	-8.84	14 83	2 128	126	-	-9-0-12-0-26-26-10-12-20-26	8.84
14102	101108-0	4011192	10	201.00102011	-10 10 244 (0.4	1.14	- 10.00	32	14.1	1.10	1.54	12	28-M-L - 8-M - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 -	1.00
14-101	1111111	101120	18.	2014/22016	120-242 (20-25) #	2.25	111	금란	11 11	1.12	10	12	관음학학교공공급적	1 22
18212 0	1411424	3114268	6.26	a 161 mil Mil 4	. ALEDIMID #	16.24	(6.11		218 24	1 218	1.24	1.0	48-8-18-18-17-17-28-28-88-81-5	1 84
18.211	1111-08-01	1111005	521.1	261.ev+141.3	+10.001 8.84 MI	141	-b41	-3.84	28.64	1. 143	245	40.	-8-8-848484848484848484848	8.44
112	21120	네말랐	10.1	121202	10.00 100.00	-10	.74	ゴ幕	12 2	1 12	12	12	3332333373	1.10
14.115	101102.00	10.01.001	11	 201.008 801 1 201.008 104 0 	10.00110.001	1 10	241		101 101	1.1	1.00	12	21 - 11 - 11 - 11 - 1 - 12 - 12 - 13 - 14 - 1 - 12 - 12 - 12 - 12 - 12 - 12	1 83
18.215	1411-0023	-8124 (8.5	141	at works in	+18.022244.00	1-04	29-61	-62.24	10 51	2 1.18	1.65	82	19-10-1-0-1-0-10-10-10-10-10	8.19
1820	#5111206	10000	100	 201.003 Kbi in 201.003 Kbi iii 	200 THE 200 MIL II	1.2	246	-14	10 10	1.10	805	18	all and a faile of all all all all all all all all all al	1.00
18.2-51	14111245	1000	12	 241.000 241.4 241.000 241.4 	100 100 100 cm at		14	-5.84	11 11	1.14	45	10	18-8-8-18-5-1-8-1-18-88-8	1.44
18.171	851153.54	- 14 14 184	18	201.022.201.0	118-042 (404)	1.14	244	-11.14	100.00	1 14	100	10	off a Contact of all all a builded	2.00
D&LED	141112.14	-8114 10.0	14.15	261.672.181.6	-18-01020101	1.41	-16.44	+61.24	112 14	1.16	1.24	1.04	28-81-14-86-04-11-0-86-0-08	2 23
1211	21122	11031	12	12122213	-2.1212.5	122	38	.12	11 11	: 23	12	12	**************	1:33
DATE:	NUTCHAN	-10.00	- 20	• 262 mm hds m	- 10 101 101 101	1.74	-20.64	-1124	1.4 .04	5 1.4	1.84	1.0	21 - 2 -11 - 0 -21 -0. x 0 -18 x 0 - 4	2 24
18412	141112.45	14210-012	100	201-000-000-000-0	+20 50 244 30	156	111	- 102.84	1.3 1.1	114	11	15	all all a ball all all a state of a second	2 22
10.110	PA 12 Martin	1002104	10	 246.02712410 266.02712610 	100 ATM 758.071	1.53	1.00	-1.12	201 10	1.55	121	12	- 1 1 H - 1 2 K 1 4 2 H H H - 1 1 H	1 11
IA.L III	111110-00	1015.000	684	206-011261-2	107-610 Mail 01	1.01	-14	.630	118 104	1.14	1.46	18	48-8-28-28-2111-8-28-28-4	1 84
MALP'S #	10111-0024	10100	121	 Design and the 	A 10-50 MARIE I	1.15	-1-84	-1144	124. 61	1.1	1.00	10	a ball a ball of a ball of a ball	1 14
IN CO.	15110816	-21.00.26.6	1.40	201.00 Mill 1	-25 748 424 (21	-201	-141	-8.14	14 11	1.00	8.54	1ê	-9-9-10-11-10-1	1.44
CALLS:	PG 1.1 18.11	10110202	10.1	 264.020.268 264.020.266 	- NO-676 206 201	1.00	- 242	-5.26	1.0 10	14	111	100	1 P. M. 2010 121 24 27 20 20 20 20 20 2	8 -12
OR I III II	1911/10/20	-8111-6.1	181	200.021101.0	100-000 216 (C) #	10.14	-38.0	435.14		1.00	10	10	-1-84.41.49.41.1.5.1.5.14	1.14
GALLY .	14 13 18 14	100000	10	 266-6211-266-3 266-6211-266-3 	1001024440	1.41	6.21	8.23	10.00	:	101	10	20.0.10.0.1.0.0.0.0.0.0000	1.00
14.111	1413 18.14	310.027	12	161 276 261 8	-10.041.044.00	1.65	1.44	-31.24	14 44	14	111	14	10010100010000000000	1.10
10.141			100	1 14 20 40 4	1.10.000.000.00			100.00	1.2. 14	1.1	100	100	A DALL AND ALL ADDRESS	1.14
18342	151418-34	-1114-0-1	434	294-616 200 5	+10.502346422	141	.144	-214	100 00	1 10	101	15	- 5+ 8-11+0+11+8+10-18+2+3	1.00
18.245	P6 13 8008	10 14 124	3K	C 266.000 200 3	10.00 0000		6.43	-0.8		1.11	1.05		- P-B-4 - B-1 - B-1 - B-1 - B-1 - B-1 - B-21	2 1.0
14141	NA NA HEGO	-102.000	+44	6 266 (Ref) #61 (h 74	- 16-14	-41.50	14 15	6 1.68	1.04	141	a balla da ba da ba balla da ba	2 24
18.142	8513 10.05	10.06.000	10.1	 266.008.104.0 266.008.667.0 	-10.401144.76	20.00	26.16	. 10.14	112 10	1.08	121	12	- 6 - 6 - 7 - 6 - 7 - 8 - 6 - 7 - 8 - 6 - 10 - 14	2 20
18.141	8613-824	-6/14/16.8	4.84	244.00x34413	_K2-340 K34 K3	246	-2-24	-0.85	144 14	2 1.26	124	12	41-141-8-8-14-8-8-11-8-19-19	1.00
101 -	20.25	10111	18	12222	100010312	12	1.420		122.23	100	10	12	1.0.0.0.0.0.0.0.0.0.0.0.0	1.53
18301	1414-1621	18726188	4.84	 266 disk and 4 	1.02 ETN 244 C1		1.64		101 64	h #3	244	89	a balla balla da'ila dalla balla	414
18.70	1511-536		÷.	244.008.8214 214.029.8016	10.00110.00	1.00	1.00		10 10	1.1	1.14	12	2912-04-02-11-01-02-02-02-02-02-02-02-02-02-02-02-02-02-	1.11
18,205 #	1011-0.00	10110-008	10.1	 264.022 Million 264.022 Million 	100.00000000	1.1	2.41	- 10.14	1.4 1.4	1 1 1	100	10	We have been a set of a barrent of	1 10
14,203	1414 1925	-81.56 (84)	14.84	2 244 (BHD 267 H	-1649177195	2.04	-141	.216	1.0 .04	1.12	1.00	1.6	A REAL PROPERTY OF A REAL PROPERTY OF	2.00
112	111 101	생산관문	12	[엽았け]	· 建設協会	1.21	.515	글렴	12 22	2.13	111	12	<u>상품가 클럽 않던 분</u> 명권	1.23
14.701	151321.08	11111122	44	 264.000 tel 6 	10.442 814 48		-26.85	- 95.24	14 14	1.00	1.86	1.9	- 2-84-62-62-22-74-224	2 16
10.001	101226.00	10110-001	10.1	 246 (0)(247) 0 246 (0)(247) 1 	120402144.00	1.55	-144	-149	12 14	- 10	100	12	- A - C - d L - 1 - d L - 1 - d L - b - 1 - 4	1.00
18.162	15112526	10110-002	4.16	 244 dire 244 di 	12010138-00		-6.95	-2.84	121 101	2 1.28	k thd	1.0	48-8-1-9-1-1-2-8-8-8	8.40
18385 1	15112585	-1021263	15.02	a 194 dim bes di	-18.400 104 10	1.11	-3511	-2144	500 14	1 14	144	1.20	48-11 der Ball 48 a 8 a 8 a 8 a	2 82
18.105	10112-0012		18	200.000.000	1.0.02.00.0	2.65	2.05	- 10.84	1.1. 10		1.00		281521.00.01.00.00.00.00.00	8.80
ials? +	10122626		12	2 244 101 101 101 10	10.0016.014	16.61	-18.04	10.00	20.53	1.4	1.14	1.00	28 - 8 - 18 - 2 - 28 - 28 - 7 - 18 - 28	1.00
14.100	8126	-31W	12	132 귀성?	-54m1018	192		-백천	12 2	t it.	121	12	생생님생님생생생님병	1.12
ia/hi	10122-00	.8493	18	200.000.200.0	-35.011.0.00	10.04	-246	-0 M	and ma	1.12	100	12	48.8.21.01.121.121.11.11	2 23
512	111224	1212162	14	22,72523	12:00 12:00	18	154	-35#	13 11	1.12	22	1.2	세계카이(()에()()()	1 34
iain	19112-002	-11 16 18.1	16.	264 110 804 3	+18-001 200/01	1.0	1.15	***	218.00	1 14	104	10	41-81-8149-128-814118-8	1.44
ia Chi	15122636	10120100	10.1	264-001867	120.808104.00	2.00	-1.00	-110	10.00		212			5 10
ia/la	1911210-00	18716-007	18	244 CONTRACT	10.00104100.001	10-14	18.65	40.10	ad as	1.00	10	24	-1-8-8-8-8-8-8-8-8-8-8-8-8	1.24
MATCH #	P51318.85	LIN 16 (E.2	16	 264, 963 868 6 264, 963 864 6 	-10-60 101 00 m	1.45	-246	-11.84	100.10	1.12	1.14	14	- 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1	1 13
14-131	1414 (#28	-819.83	28	204 205 261 2	-19-102 146 86	8.26	10.63	-949	10 5	1.1	1.15	12	1 - 1 - 1 - 1 - 2 - 1 - 4 - 4	2.25
10.111	1514 1814	-10.14 (8.5	4.00	204-002201-0	+18-100 FM -02	1.01	-10.14	6.34	10 10	1.0	1.14	10	distaiche han die Oakstand	2.00
16.112	1513 1825	111110-004	1.40	264-0012411	128-005 KBC-01	1.14	18-01	6.24	14 14	5 int	242	iris.	all all day in a star to both all	1.49
18,211	11120	12102	10.1	122125	- 10-200 121 10	122	-13	-38	12 1	173	131	12.	100000000000000000000000000000000000000	1:15
18.205	1012-0006	-87 16 26 2	58.1	204 100 100 0	-1048420	246	0.65	28	8.1 6.0		801	8.6	- P. N. 20. 0 1N - 1 10 - 10 10 - 11	2.00
110 a	2122	110123	221	「知る間の	1:28:122。	1.12	1.44	-21B	175 23	13	12	75	243341341383	1.72
14.101	15113-0516	10124-001	440	C 264 C 8 ADI U 264 C 9 Hours	10.80 MORE	1.246	2.04	-114	121 10	1.11	114	17	- 1 - 7 - 1 all a 1 - 1 a 8 all -8 a 1 -9 - 9 - 18 -8 and - 1 - 18 -8 all -9 -9	2 .112
10.111	NUME	11140	10	264 402 104 1		2.61	1616	6.16	120 14	1.12	1.00	12	48-11-1-K-1446-PL-12-12-1	6 18
18.201	PS 12 -00.30	-8112 18.8	124	 264 (00) 201 (0 264 (00) 754 (0 	-10.001 (44.0)	100	12.41	- 101.05	118.00	1.14	111	12	48 - 80 - 20 - 80 - 80 - 70 - 10 - 80 - 80 - 80 - 80 - 80 - 80 - 8	8 24
18.251	1512-0811	-1114-042	141	284 193 844 6	-10-046 FM 20	2.16	-4-85		1.8 10	1.13	126	10	28.0.21.0.31.0.31.0.5.0.0	2 26
14.201	111.414	-212 Pe	12	212223	-21514155	12	-101	-944	11 11	1.12	1.04	12	1.1.1.1.1.1.1.1.1.1.1.1.1.1.1	111
10.100	1111-0-10	10.00	14.14	284-102-144-0	+10.010 144 76	44	1041	100.24	128 118	1.14	121	240	all - H - I all the citrate data had	8.42
MARC 1	15 13 18 18	1000	16	 244 milli Mill II 244 milli Mill II 244 milli Mill II 	-26.80°244 G	1.55	10.04	-95.84	1.4	1.18	144	12		8.43
and in	111100	-1941	10	124 40 14 3		1.55	24	1.44	12.2	: 12	11	12	-a.o.30303030303533	1.00
- and		10015184	- 48		and 178.08					- 1.8		-5		

into...

The Milky Way from Gaia

Image credit: NASA/JPL-Caltech/R. Hurt (SSC)

Gaia Catalogue complexities Modelling our Galaxy Wish list for data archive

La Palma 30.05.2011 - p.28/47

Determining the best Galaxy model

Goal

• Dynamical model of the Galaxy capable of 'explaining' the entire Gaia catalogue

Dynamical model

- gravitational potential
- distribution functions for each stellar population
 - probability distributions in mass, abundances, ages
- large number of parameters

Basic predictions:

 $f_{\Theta}(\mathbf{r}, \mathbf{v}, t | \text{stellar population})$

- 1. In what space do we work to determine the parameter vector Θ ?
- **2**. *How* do we determine the parameter vector Θ ?

Direct comparison in phase-space

Directly comparing predicted (\mathbf{r}, \mathbf{v}) to observed values is most 'natural' approach. However:

- Effects of dust to be corrected
- Incomplete phase space data (lack of v_{rad})
- Parallax to distance conversion is non-linear $(r = 1/\varpi)$
 - For true and observed parallax ϖ_0 and ϖ :

$$E[\varpi] = \varpi_0 \quad \text{but} \quad E[\frac{1}{\varpi}] \neq \frac{1}{\varpi_0} \,!$$

- For large σ_{ϖ}/ϖ expect large biases and spurious features in **r**, **v** *E*, *L*, etc
 - *E* and *L* both depend on $1/\varpi^2$
- Selection on σ_{ϖ}/ϖ can introduce severe truncation biases
- Correlated observables and non-linear transformation of parallax will produce strongly non-Gaussian errors with complicated correlations

Direct comparison in phase-space

See: Brown, Velázquez & Aguilar, 2005, arXiv:astro-ph/0504243

Direct comparison in phase-space

See: Brown, Velázquez & Aguilar, 2005, arXiv:astro-ph/0504243

Forward modelling

Project Galaxy model into the data space:

```
\begin{array}{l} f_{\Theta}(\mathbf{r},\mathbf{v},t|\text{stellar population}) \rightarrow \\ g_{\Theta}(\alpha,\delta,\varpi,\mu_{\alpha*},\mu_{\delta},\nu_{\text{rad}},G,\text{colour},A_V,[\text{M}/\text{H}],\dots|\text{stellar population}) \end{array}
```

- No non-linear parallax transformation
- All parallax data can be used (including negative parallaxes)
- Easy accounting for incomplete phase space data (i.e., no v_{rad})
- Selection effects and incompleteness can be modelled
- Extinction can be modelled
- Correlations in the measurement errors are more easily accounted for in the data space

'Deciding' on the best parameter values

'Deciding' on the best parameter values

Bayesian inference through maximization of likelihood:

$$\phi = \sum_{\text{all stars}} \ln p(\alpha, \delta, \varpi, \mu_{\alpha*}, \mu_{\delta}, \nu_{\text{rad}}, G, \text{colour}, A_V, [M/H], \dots | \text{stellar population}, \Theta)$$

- Allows precise hypothesis testing
- Result will be a probability distribution over the parameter space
 - not all aspects of Galaxy model will be uniquely determined

Challenges:

- Large amount of data and large number of model parameters
- Full catalogue comparison needed
 - constraints from all sky directions and full Gaia volume
 - Star-to-star correlations to be accounted for

Facilitating precise hypothesis testing

Three proposals in Hogg & Lang (2011, arXiv:1008.0738)

- Careful definition of the uncertainty of each catalogue entry
- Provide a sampling of possible Gaia catalogues
 - approximate the full covariance matrix of the catalogue
- Expose the likelihood function
 - provide access to the full covariance matrix

Definition of catalogue entry uncertainty

- Catalogue entry $\mathbf{v}^T = (\alpha, \delta, \varpi, \mu_{\alpha*}, \mu_{\delta}, \nu_{rad}), \mathbf{C}^{-1}$
- Two hypotheses (alternatives) for \mathbf{v} ; \mathbf{V}_1 and \mathbf{V}_2

define

$$\Delta \chi^2 \equiv (\mathbf{V}_2 - \mathbf{v})^T \mathbf{C}^{-1} (\mathbf{V}_2 - \mathbf{v})^T - (\mathbf{V}_1 - \mathbf{v})^T \mathbf{C}^{-1} (\mathbf{V}_1 - \mathbf{v})^T$$

- Define v, \mathbf{C}^{-1} such that the result is as close as possible to what you would have computed for $\Delta \chi^2$ in the *raw image pixels*, marginalizing over all nuisance (calibration) parameters.
- For practical example see 'Spectroperfectionism': Bolton & Schlegel, arXiv:0911.2689
- Not far from existing plans for publication of Gaia data (see also Hipparcos Catalogue)
- How to do this transparently for the great variety of catalogue entries?
 - differences in 'distance to the raw data'

Make K + 1 Gaia catalogues

- Release one 'primary' catalogue
- Other *K* catalogues sample posterior distribution of possible catalogues
 - astrometric, radial velocity, and astrophysical parameter variations
 - calibration parameter variations
 - allow catalogues of different complexity (eg., binary vs. single star)
- Average of a quantity over the *K* samples should approximate marginalization over all probabilistic quantities
- Approximates the full covariance matrix of the catalogue
 - star-to-star correlations can be calculated
- For a practical example see presentation by Berry Holl
- How large should *K* be?
- How to sample 'catalogue space' efficiently such that the result is representative of the posterior distribution?
 - allowing for complexity variations is challenging

- Allow users to submit a Δ catalogue
 - difference between primary Gaia catalogue and an alternative
 - > Δ can be for science data and/or calibration parameters
- Return $\Delta \log \mathcal{L}$
 - likelihood re-evaluated against the raw pixel data
 - provides access to any covariance matrix element
 - computational burden on the user (performs his/her own uncertainty analysis)
- Allows handling different complexity choices
 - alternative deblending in crowded regions
 - single vs. multiple star solutions
 - incorporation different priors
 - different astrophysical parameters from improved atmosphere models
 - ...

But this is crazy...

- Allow users to submit a Δ catalogue
 - difference between primary Gaia catalogue and an alternative
 - > Δ can be for science data and/or calibration parameters
- Return $\Delta \log \mathcal{L}$
 - likelihood re-evaluated against the raw pixel data
 - provides access to any covariance matrix element
 - computational burden on the user (performs his/her own uncertainty analysis)
- Allows handling different complexity choices
 - alternative deblending in crowded regions
 - single vs. multiple star solutions
 - incorporation different priors
 - different astrophysical parameters from improved atmosphere models.

66

(Parts of) Gaia data processing chain already use forward modelling:

(Parts of) Gaia data processing chain already use forward modelling:

Why not publish this machinery? Think 2020 and beyond!

Preserve the raw data and processing software

Data curation

- All raw data
 - $\sim 60 \text{ TB}$ uncompressed
- Calibration data and models
- Intermediate data products
- All processing software
- Implement data lineage concept

Science goals

- Raw data reprocessing based on better algorithms, better calibration models etc
- Alternative processing of specific stars, groups of stars, or even entire catalogue
- Reprocessing data based on new and independent information

Database and data archive

- General public oriented with advanced features for professionals
- Allow arbitrary queries
 - with option to account for star-to-star covariances
- Data mining
- Visualisation (data space is > 10-dimensional!)
- Transparent integration with other catalogues or sky surveys
- Access with variety of 'devices'
- All of this has to be fast of course ...

Living archive concept

Gaia catalogue and archive released in ~ 2021 should not be 'final'

Photo Courtesy of the Isaac Newton Group of Telescopes, La Palma

- Updates should be allowed so as to incorporate:
 - updated classification or parametrization of stars
 - better distance estimates for faint stars
 - ground-based follow-up observations
 - independent information on, e.g., double stars
- Implications for maintenance, quality, security, keeping mirrors in sync

Bring the processing to the data

- Szalay (Sloan Digital Sky Survey) has advocated this for large archives
 - allow arbitrarily complex processing of archive data
 - example: dynamical model of the Milky Way that best explains the catalogue
- Virtualisation (O'Mullane, ESAC) could allow users a virtual machine *in the Data Centre with the Archive*
 - code what you want and specify how you want run it

Let's work to make the Gaia archive future proof!

- Gaia will provide an unprecedented stereoscopic map of our Milky Way and the nearby universe
 - I billion stars, 300 000 solar system objects, millions of galaxies, 500 000 quasars, 10 000 exo-planets, ...
 - catalogue 'finished' in 2021
- It will be *the* astronomical data archive for decades to come
 - tremendous discovery potential when combined with other archives
- Research or invest effort in the following:
 - keep raw data, calibration data, and processing software available
 - facilitate reprocessing
 - make the archive 'live'
 - bring the processing to the data
 - precise hypothesis testing agains the raw image pixels