LSST Data Management: Summer 2013 Data Challenge Plan

LSST Data Management

Friday 19th April, 2013, 00:18

1 Plans

1.1 Object Characterization

In general, LSST/DM needs to be able to fit a linear combination of analytic models to observed objects. The obvious ones we need to support are elliptical Sersic and point source models. The former are crucial for extended source characterization and shape measurement and will be the focus of pre-FDR development.

In the near term (in pre-FDR development phase), we plan to implement code that will fit a linear combination of bulge (Sersic $n = 4$) and exponential disk (Sersic $n = 1$) models with common ellipticity. The free parameters include: ellipticity (e_1, e_2), scale lengths (l_1, l_2), and the bulge and disk amplitudes, a_1, a_2 (for a total of 6 parameters). For now, we will keep the bulge-to-disk scale length ratio fixed; this is the same approach that CFHTLens collaboration is taking (Miller et al. 2013). That leaves us with a 5-parameter fit.

Ultimately, we will also need to let the centroids vary, for a total of 8 (or 7) parameters. However, as we're mostly interested in marginalizing over the position, we are allowed to be clever about how to do this. This will be left for post-FDR development, as will the point-source model fits (that allow for proper motions).

Pre-FDR, we plan to implement a fitter that, given a series of postage stamps of an extended object, will fit for the five free parameters (as described above). The fitter can be thought of as consisting of two parts: a) the model evaluation code, which computes the pixel values of the model given the parameters and b) an optimizer, which (iteratively) compares the pixels computed by the model evaluation code to real data and determines the posterior distribution of parameters given the data and the applicable prior.

The model evaluation code will use the fact that Sersic models can be decomposed to a mixture of gaussians (see Bosch 2010, Hogg Lang 2012), to quickly convolve the models with the PSF. For this to work, the PSF will be decomposed to a shapelet basis. The optimizer will perform a non-linear fit of ellipticities and overall scale (the scale ratio will initially be kept fixed). We plan to use importance sampling and therefore the analysis of expected shapes of likelihood surfaces will be necessary to determine a good proposal distribution to sample from. Given the ellipticities and scales, the fit of amplitudes is linear and this will be exploited to speed up the optimization. One area of research to be done pre-FDR is to understand how to incorporate priors on amplitude (e.g., on bulge-disk ratios) in a way that will not compromise this speed-up.

Generically, the output of this fit will be a series of independent samples from the posterior distribution of model parameters given the data. In the long-run, we will investigate how many samples per galaxy are needed and what is the best way to encode and store this information.

We will begin by examining fits of a single Sersic model to single epoch data, then move on to understanding double-Sersic models (note: double-Sersic fit code will be implemented right away, but will allow zeroing out the amplitude of the second). A single epoch fit requires both the model evaluation code, and the optimizer code. To write the importance sampling optimizer we need some understanding of what the likelihood surfaces look like, so we'll start with a brute-force optimizer (evaluation on a dense grid). Those results will be visualized and a proposal distribution chosen for the importance sampler.

We will look into the selection of a reasonable prior for Sersic model parameters. This prior will also be needed to inform the image simulations needed to test the object characterization code. The most important priors to set are on bulge and disk scale lengths (the ratio will be fixed, but we need a prior on the overall radius, and the value at to which to fix the ratio). We may also need a prior on ellipticity (which can be deduced from looking at ImSim data). We do not expect to need a prior on total flux, as we hope it to be well constrained by the likelihood.

The importance sampler will then be implemented, and its speed and accuracy assessed (the latter by comparison with ground truth). All scientific and computational performance benchmarks will be performed against ImSim-simulated data. We will simulate galaxies in a range of seeings, S/N ratios, and with a range of input model parameters to understand its performance and sketch out the needs for further (post-FDR) development.

In parallel, we plan to complete CoaddPsf (the LSST implementation of StackFit-computed PSF; Jee et al. 2011), integration of meas_mosaic (the relative astrometry package enabling precise registration of multiple epochs going into co-adds), and work on improving the quality of shape measurement on coadds of simulations (using the shapeHSM measurement code until the new modeling code is available).

Finally, the code will be exercised in multifit mode (with multiple epochs). Speed and accuracy will be evaluated. We will measure the performance improvement obtained by using a StackFit-provided proposal distribution.

To summarize, the object-characterization related goals for LSST/DM pre-FDR phase are to:

1. Complete the tuning and testing of existing code to build coadds and measure the shapes of galaxies on those using CoaddPsf (aka. StackFit), first with the existing shapeHSM package, and eventually with double-Sersic modeling code.

2. Write and test the code that will fit (in the sense of sampling the posterior) 5-parameter double-Sersic models to multi-epoch images of galaxies (aka. MultiFit);

3. Assess its performance and confirm that we're on the path to achieve the performance necessary for operations. The ultimate goal is to show that MultiFit object characterization can be performed in under $\sim$3 seconds per object per epoch.

There is no requirement for the object characterization code produced in pre-FDR phase to work on real data; we plan to do all performance assessment on simulations. Secondly, PSF characterization is out of scope for pre-FDR; we will assume it will be characterized reasonably well once this code is written (post-FDR), and use ImSim to emulate our expected knowledge of the PSF.

1.2 Middleware Scaling and Split Production Tests

This goal is currently documented in PMCS only.

1.3 Continued Image Differencing Develoment

This goal is currently documented in PMCS only.

2 Report

This section will be written upon completion of Summer 2013 Data Challenge.

References

[1] Miller et al 2013,http://adsabs.harvard.edu/abs/2013MNRAS.429.2858M

[2] Hogg & Lang 2012, http://arxiv.org/abs/1210.6563

[3] Bosch 2010, http://arxiv.org/abs/1007.1681

[4] Jee & Tyson, http://adsabs.harvard.edu/abs/2011PASP..123..596J