Evaluation of Estimation Accuracy without a Gold Standard

Imaging systems are often used for the purpose of estimating a quantity of interest that is indicative of a patient's health status. Often, this quantity of interest can be computed on a variety of imaging modalities or using a variety of image-processing techniques. For example, cardiac ejection fraction (EF) is a measure of the fraction of blood pumped out of the heart at the end of each heart cycle. EF can be assessed by a variety of imaging modalities, including nuclear medicine, ultrasound, X-ray angiography, and MRI. The estimates from all of these modalities may differ, and none of them can be considered the "gold standard". Thus, one is left to compare the estimates from one modality to those of another, which does not indicate how either method actually relates to the true gold standard. We have developed a technique to assess how estimates of EF from different modalities relate to the gold standard without knowing what the gold standard actually is. This technique was developed in 2003 and studied extensively in simulation studies. Only recently, through collaboration with researchers at the University of Massachusetts Medical School, have we been able to apply this technique to real patient cardiac EF measurements computed using three different segmentation algorithms. In completing this work, we were forced to address an issue that has plagued the technique since its inception: With real data, how will we know when the technique is working when we will not have access to the gold standard? To address this issue, we have developed three *consistency checks*. If the results of the technique fail any of these consistency checks, then we are very confident that the model used or the estimates returned by the method are not useable. However, if the results pass all three consistency checks, then that garners confidence in both the model and the estimates returned by the method. That is, failure of a check implies that the method failed; success on all three checks doesn't necessarily imply that the method worked.

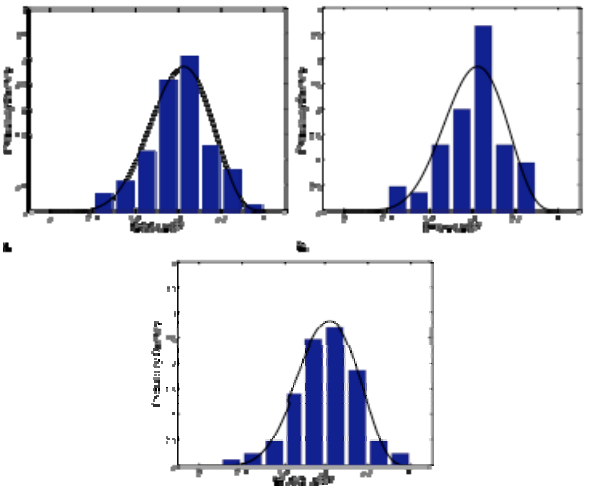

Fig. 1. An example of one of the consistency checks. Notice that the estimated distribution of cardiac EFs matches that of the raw data adjusted by the parameters returned by the method.

M. A. Kupinski, J. W. Hoppin, J. Krasnow, S. Dahlberg, J. A. Leppo, M. A. King, E. Clarkson, and H. H. Barrett, Comparing cardiac ejection fraction estimation algorithms without a gold standard, *Acad. Radiol.*, 31:329-337, 2006.

J. W. Hoppin, M. A. Kupinski, G. A. Kastis, E. Clarkson, and H. H. Barrett, Objective comparison of quantitative imaging modalities without the use of a gold standard, *IEEE Trans. Med. Imag.*, 21, 5:441-449, 2002.