layer to be split, the first move works by randomly selecting two seed super-pixels and then assigning all remaining super-pixels to the closest (in appearance space) seed. The initial seeds are chosen such that with high probability they are far in appearance space. The second move employs a connected component operation. If the given layer has disconnected components then one such disconnected component is sampled at random and deemed to be a new layer.

Swap. The swap move reorders the layers in the current partition, by selecting two layers and exchanging their order.

Shift. The shift move refines the partitions found by the other moves. It iterates over all super-pixels in the image assigning each to a segment which maximizes the posterior probability [4]. Observe that the merge and split moves change the number of layers in a partition performing model selection, while swap and shift attempt to find the optimal partition given a model order.

4. Learning from Human Segmentations

In this section, we provide methods for quantitatively calibrating the proposed models to appropriate human segmentation biases. Recall that our model has four hyper-parameters, the PY region size hyper-parameter (α), the appearance hyper-parameter (ρ) and the GP covariance parameters (A and Ψ). We tune these to the human segmentations from the 200 training images of the Berkeley Segmentation Dataset (BSDS) [12]. We show that in spite of the inherent uncertainty in the segmentations of an image, we are able to learn important low level grouping cues.

Learning size and appearance hyper-parameters. The optimal region size hyper-parameters are the ones that best describe the statistics of the training data. We select $\hat{\alpha} = (\hat{\alpha}_a, \hat{\alpha}_b)$ by performing a grid search over 20 evenly spaced α_a and α_b candidates in the intervals $[0, 1]$ and $[0.5, 20]$ respectively and choosing values which maximize the model's likelihood of the training partitions according to equation 8. The appearance hyper-parameters $\hat{\rho} = (\hat{\rho}^t, \hat{\rho}^c)$ are tuned through cross validation on a subset of the training set. For BSDS, the estimated parameters equal $\hat{\alpha}_a = 0.15, \hat{\alpha}_b = 1$ $\hat{\rho}^t = 0.01$ and $\hat{\rho}^c = 0.01$

Learning covariance kernel hyper-parameters. The covariance kernel governs the type of layers that can be expressed by the model. Estimating it accurately is crucial for accurately partitioning images. In [25, 24] the authors use various heuristics to specify this kernel. Here, we take a more data driven approach and learn the kernel from human segmentations. While we cannot expect our training data to provide examples of all important region appearance patterns, it does provide important cues. In particular like [9], we learn to predict the probability that *pairs* of super-pixels occupy the same segment via human segmentations.

For every pair of super-pixels, we consider several potentially informative low-level cues: (i) pairwise Euclidean distance between super-pixel centers; (ii) intervening contours, quantified as the maximal response of the probability of boundary (Pb) detector [13] on the straight line linking super-pixel centers; (iii) local feature differences, estimated via log empirical likelihood ratios of χ^2 distances between super-pixel color and texture histograms [20]. To model non-linear relationships between these four raw features and super-pixel groupings, each feature is represented via the activation of 20 radial basis functions, with the appropriate bandwidth chosen by cross-validation. Concatenating these gives a feature vector ϕ_{ij} for every super-pixel pair i, j. We then train a L_2 regularized logistic regression model to predict the probability of two super-pixels occupying the same segment q_{ij}. Figure 2 illustrates the effect of these cues on partitions preferred by the model.

When probabilities are chosen to depend only on the distance between super-pixels the distribution constructed defines a generative model of image features. When these probabilities also incorporate contour cues, the model becomes a conditionally specified distribution on image partitions, analogous to a conditional random field [10].

From probabilities to correlations. Recall that our layers are functions sampled from multivariate Gaussian distributions, with covariance Σ with unit variance and a potentially different correlation c_{ij} for each super-pixel pair i, j. For each super-pixel pair, q_{ij} is *independently* determined by the corresponding correlation coefficient c_{ij}. As detailed in the supplement there exists an one-to-one mapping between the pairwise probabilities and correlations, allowing us to go from the logistic regression outputs (q_{ij}) to correlation matrices. These correlation matrices (C), learned from pairwise probabilities will in general not be positive semi-definite (PSD). We cope by finding the closest PSD unit diagonal matrix to the correlation matrix. We use the recently proposed technique of Borsdorf *et al.* [3], which solves for A and Ψ by minimizing the Frobenius norm$||C - (AA^T + \Psi)||_F$. It should be noted that even the heuristic approaches of Sudderth and Jordan [25] and Shyr *et al.* [24] can yield non PSD correlation matrices. There the authors ensure positive semi-definiteness by performing an eigen-decomposition of C and retaining only non-negative eigenvalues. This is a cruder approximation and leads to poor results (Figure 2).

5. Spatially dependent PY model properties

In this section, we explore various properties of our model which may not be immediately obvious.

[4] A naive shift move would evaluate the posterior probability of the partition after every super-pixel shift. This proves to be prohibitively expensive, instead we develop an alternative which allows us to evaluate the posterior after one complete sweep through the super-pixels while ensuring that each individual shift by-and-large increases the posterior. Please see the supplement for details.