Nonparametric Bayesian Line Detection
Towards Proper Priors for Robotic Computer Vision

Anne C. van Rossum[1,3], Hai Xiang Lin[2,3], Johan Dubbeldam[2] and H. Jaap van den Herik[3]
[1]*Almende B. V. and Distributed Organisms B. V. (DoBots), Rotterdam, The Netherlands*
[2]*Delft University of Technology, Delft, The Netherlands*
[3]*Leiden University, Leiden, The Netherlands*

Keywords: Bayesian Nonparametrics, Line Detection.

Abstract: In computer vision there are many sophisticated methods to perform inference over multiple lines, however they are quite ad-hoc. In this paper a fully Bayesian approach is used to fit multiple lines to a point cloud simultaneously. Our model extends a linear Bayesian regression model to an infinite mixture model and uses a Dirichlet process as a prior for the partition. We perform Gibbs sampling over non-unique parameters as well as over clusters to fit lines of a fixed length, a variety of orientations, and a variable number of data points. The performance is measured using the Rand Index, the Adjusted Rand Index, and two other clustering performance indicators. This paper is mainly meant to demonstrate that general Bayesian methods can be used for line estimation. Bayesian methods, namely, given a model and noise, perform optimal inference over the data. Moreover, rather than only demonstrating the concept as such, the first results are promising with respect to the described clustering performance indicators. Further research is required to extend the method to inference over multiple line segments and multiple volumetric objects that will need to be built on the mathematical foundation that has been laid down in this paper.

1 INTRODUCTION

In computer vision and particularly in robotics, traditionally the task of line detection has been performed through sophisticated, but ad-hoc methods. We will give two examples of such methods. RANSAC (Bolles and Fischler, 1981) is a method that iteratively tests a hypothesis. A line is fitted through a subset of points. Then other points that are in consensus with this line (according to a certain loss function) are added to the subset. This procedure is repeated till a certain performance level is obtained. The Hough transform (Hough, 1962) is a deterministic approach which maps points in the image space to curves in the so-called Hough space of slopes and intercepts. A line is extracted by getting the maximum in the Hough space.

There are four main problems with these methods. First, the extension of RANSAC or Hough to the detection of multiple lines is nontrivial (Zhang and Ksecká, 2007; Gallo et al., 2011; Chen et al., 2001). Second, the noise level is hardcoded into model parameters and it is not possible to incorporate knowledge about the nature of the noise. Third, it is hard to extend the model to hierarchical forms, for example, to lines that form more complicated structures such as squares or volumetric forms. Fourth, there are no results known with respect to any form of optimality of the mentioned algorithms.

Bayesian methods (Fienberg et al., 2006) are nowadays commonplace to solve ill-posed problems. A problem is defined by a likelihood function and by postulating a prior. Bayes rule subsequently gives the unique, optimal solution of combining the likelihood with the prior to obtain the posterior from the viewpoint of information processing (Zellner, 1988). Note, that optimality about the inference procedure does not say anything about the correctness of the likelihood function or the postulated prior.

The detection task of multiple lines might seem a rather straightforward problem, but a proper definition will be useful for many application domains. In robotics depth sensors generate large point clouds of data that are difficult to process in its raw form. Compression of this data into lines, planes, and volumetric objects (Kwon et al., 2004) is of paramount importance to accelerate the inference in, for example, simultaneous localization and mapping (Vasude-

Rossum, A., Lin, H., Dubbeldam, J. and Herik, H.
Nonparametric Bayesian Line Detection - Towards Proper Priors for Robotic Computer Vision.
DOI: 10.5220/0005673301190127
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 119-127
ISBN: 978-989-758-173-1

van et al., 2007). A method that is able to infer multiple lines simultaneously can be extended to perform inference over multiple planes and objects. Moreover, the Bayesian approach will allow for setting intriguing priors, that for example introduce a prevalence for certain horizontal and vertical angles in man-made environments compared to more natural scenes (as seems to the case for the number of unique objects (Sudderth and Jordan, 2009).

In this paper we will postulate a method to perform inference over the number of lines and over the fitting of points on that line. To achieve this, we require methods from the field of Bayesian nonparameterics. Probabilistic and even Bayesian extensions to the Hough transform exist (Bonci et al., 2005; Dahyot, 2009), but until now researchers have not been separating the model used to infer an individual line from the model to infer a number of them.

2 BAYESIAN NONPARAMETRICS

In machine learning there are many methods that require a predefined figure for the number of items to be recognized. The most well known example is the parameter "k" in k-means clustering which fixes the number of clusters to search for. The Bayesian approach towards a multi-object estimation problem is to provide a prior on the number of clusters that allows this number to be (in theory) from one to infinity. A naive interpretation would require an integral over an infinite number of models. This can be prevented by performing inference over partitions of the data. The data will be finite for all practical applications.

Apart from a prior over the number of partitions, there should also be a prior formulated with respect to the distribution of points over these partitions. Note, that our line detection task is actually a partition problem. We are interested in which points belong to which line and we want to know the parameters of each line. However, we have no preferred index for the lines themselves. They are neither ordered in a specific manner, nor do they have labels. This property of a partition is called exchangeability.

2.1 Dirichlet Process

The exchangeability (de Finetti, 1992) property is related to conditional independence by de Finetti's theorem. The theorem states that for exchangeable observations there is some hidden random variable that make the observations conditionally independent (and have the same joint probability distribution). De Finetti's theorem does only reveal the existence of this random variable, nothing more. In our case we will see that for infinitely exchangeable sequences the so-called Dirichlet process is a set of such random variables.

The Dirichlet process is a distribution over function spaces in which these function spaces are probability measures in their own right. Suppose we have a parameter set $\Theta = \{\theta_0, \ldots, \theta_N\}$, with θ_i corresponding to observation w_i (in our case an observation w_i exists of a tuple $\{X_i, y_i\}$), then we describe the Dirichlet process as follows:

$$G \sim DP(\alpha, H) \tag{1}$$

This means that for every (finite measurable) partition $\{A_0, \ldots, A_k\}$ of the parameter set Θ, the random distribution G is a Dirichlet process with base distribution H and concentration parameter α:

$$\{G(A_0), \ldots, G(A_k)\} \sim Dir(\alpha H(A_0), \ldots, \alpha H(A_k)) \tag{2}$$

It is important to pay close attention to indices. The Dirichlet process samples from a continuous base distribution H. However, the samples themselves can be discrete in the sense that parameter θ_j tied to observation j can be exactly the same as parameter θ_k tied to observation k.

2.2 Dirichlet Mixture Model

The Dirichlet process can be used as a mixture model (Antoniak, 1974; Escobar and West, 1995; MacEachern and Müller, 1998) in which it generates (non-unique) parameters that subsequently generate observations:

$$\begin{aligned} G &\sim DP(\alpha, H) \\ \theta_i \mid G &\sim G \\ w_i \mid \theta_i &\sim F(\theta_i) \end{aligned} \tag{3}$$

Here F describes the mapping from parameters θ_i to observations w_i. It is possible to integrate over G and sample the parameters directly from the base distribution H.

2.3 Gibbs Sampling of Parameters

Gibbs sampling requires the conditional probabilities of all entities involved (Geman and Geman, 1984). Gibbs sampling just as other Markov chain Monte Carlo methods generates a sequence of correlated samples. Subsequently, if necessary, the Maximum A Posteriori estimation of a value can be found through picking the mode (most common occurring value) of a parameter.

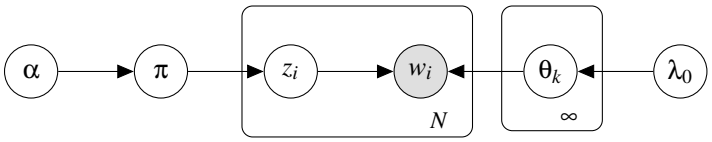

Figure 1: The Bayesian linear regression model for multiple lines in plate notation (Buntine, 1994). A nice name might be the Infinite Line Model. The Dirichlet process is defined at the left with concentration parameter α. It generates the partitions $(\pi_1, \ldots, \pi_k)$ with assignment parameters z_i that denote which observation i belongs to which cluster k. The cluster is summarized through the parameter set θ_k and has λ as its hyperparameter.

The derivation of the conditional probabilities of parameters with respect to the remaining parameters has been described in the literature (Neal, 2000). Such a derivation uses an important property of the Dirichlet process, namely that it is the conjugate prior of the multinomial distribution. Thanks to conjugacy the following equations have closed-form descriptions. The conditional probabilities are sampled from the base distribution G_0 and the other parameters θ_i in the following way:

$$\theta_{n+1} \mid \theta_1 \ldots \theta_{n-1} \sim \frac{1}{\alpha+n}(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}) \quad (4)$$

If we include the observations themselves, we need to include the likelihood as well:

$$\theta_i \mid \theta_{-i}, w_i \sim C\left\{\sum_{i \neq j} F(w_i, \theta_j)\delta_{\theta_j} + \alpha H_i \int F(w_i, \theta) dH(\theta)\right\} \quad (5)$$

The constant C is a normalization factor to make the above a proper probability density (summing to one). The entity H_i is the posterior density of θ given H as prior and y_i as observation. The notation θ_{-i} describes the set of all parameters Θ with θ_i excluded. The integral over $dH(\theta)$ is a Lebesgue-Stieltjes integral that weighs the contribution of $F(w_i, \theta)$ with the base distribution $H(\theta)$.

Equation 5 can be used to perform inference directly with all (non-unique) parameters θ_i tied to observations w_i. Details on inference will be provided in Sect. 3.

2.4 Gibbs Sampling of Clusters

It is also possible to iterate only over the clusters. The derivation takes a few steps (Neal, 2000) but leads to a simple update for the component indices that only depends on the number of data items per cluster, the parameter α, and the data at hand.

The probability to sample from a cluster depends on the number of items in that cluster (except the data item at hand). This is expressed in equation 6.

$$p(c_i = c \text{ and } c_i = c_j \text{ and } i \neq j \mid c_{-i}, w_i, \alpha, \theta) \propto \frac{n_{c,-i}}{\alpha+n-1} F(w_i \mid \theta_i) \quad (6)$$

The probability to sample a new cluster only depends on α and the total number of data items. This is described in equation 7.

$$p(c_i \in \Omega(c) \text{ and } c_i \neq c_j \text{ and } i \neq j \mid c_{-i}, \alpha) \propto \frac{\alpha}{\alpha+n-1} \int F(w_i \mid \theta_i) dH(\theta) \quad (7)$$

Here $\Omega(c)$ denotes all admitted values for c_i.

The importance of conjugacy is obvious from Eq. 7, it will lead to an analytic form of the integral. The inference method using equations 6 and 7 is described in section 3.

3 MODEL

The proposed model extends the Bayesian linear regression to multiple lines using a Dirichlet process as a prior for the partitioning of points over lines and the number of lines overall. We will name this model the "Infinite Line Mixture Model". This name follows the naming convention for other models in the nonparametric Bayesian literature (Rasmussen, 1999; Ghahramani and Griffiths, 2005; Gael et al., 2009). in particular, "infinite" means that there are an infinite number of lines to be inferred (see Figure 1).

3.1 Bayesian Linear Regression Model

Let us first reiterate the Bayesian linear regression model for a single line (Box and Tiao, 2011). A line is assumed to have Gaussian noise. For the individual points i we can write this as a Normal distribution:

$$y_i \sim \mathcal{N}(x_i\beta, \sigma^2) \quad (8)$$

The coordinate (column) vector β maps the (row) vector with independent variables x_i to the dependent variable y. The noise is normally distributed with standard deviation σ along the dimension of the dependent variable. In a computer vision task with

Algorithm 1: Gibbs sampling over parameters θ_i.

procedure GIBBS ALGORITHM 1(w, λ_0, α) ▷ Accepts points w, hyperparameters λ_0, α and returns k line coordinates
  for all $t = 1 : T$ **do**
    for all $i = 1 : N$ **do**
      for all $j = 1 : N, j \neq i$ **do**
        L_j = likelihood(w_i, θ_j) ▷ Update likelihood for all theta (except θ_i) given observation w_i
      end for
      P_i = post_pred(w_i, λ_0) ▷ Posterior predictive of w_i given hyper parameters
      $p(new) = \frac{\alpha P_i}{\alpha P_i + \sum_i L_i}$ ▷ Sample new or old?
      if $p(new)$ **then** ▷ Informal notation for sampling with probability $p(new)$
        λ_{temp} = update(w_i, λ_0) ▷ Update sufficient statistics with observation w_i
        $\theta_i \sim NIG(\lambda_{temp})$ ▷ Sample θ_i from NIG
      else
        θ_i sampled from existing clusters ▷ Sample old cluster
      end if
    end for
  end for
  return summary on θ_k for k lines
end procedure

images $x_i = [1, x_{value}]$ and y_i is the y_{value}. The x-coordinate value is transformed to obtain a value for the intersect for β_0.

All observations that belong to the same single line lead to a likelihood function that corresponds to a normally distributed random variable with y and X as parameters:

$$p(y \mid X, \beta, \sigma^2) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right) \quad (9)$$

The dependent variable is now a column vector of values y and each observation has a row of independent variables in X. The coordinate vector β and the standard deviation σ are shared across all observations.

3.2 Conjugate Prior for the Bayesian Linear Regression Model

The conjugate prior has the form of Eq. 9 which can be composed out of a separate prior for the standard deviation $p(\sigma)$ and the conditional probability of the line coefficients given the standard deviation $p(\beta \mid \sigma^2)$.

$$p(\sigma^2, \beta) = p(\sigma^2)p(\beta \mid \sigma^2) \quad (10)$$

The standard deviation σ is sampled from an Inverse-Gamma (IG) distribution:

$$p(\sigma) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{1}{2\sigma^2}\nu_0 s_0^2) \quad (11)$$

This is an $IG(a, b)$ with $a = \nu_0/2$ and $b = 1/2\nu_0 s_0^2$. The conditional with respect to the line coefficients has a normal distribution as prior:

$$p(\beta \mid \sigma^2) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0)\right) \quad (12)$$

3.2.1 Sufficient Statistics

Due to the fact that it is a conjugate distribution we have a simplified description for updating the parameters at once, given a set of observations. The sufficient statistics are updated (Minka, 2000) according to:

$$\begin{aligned} \Lambda_n &= (X^T X + \Lambda_0) \\ \mu_n &= \Lambda_n^{-1}(\Lambda_0 \mu_0 + X^T y) \\ a_n &= a_0 + n/2 \\ b_n &= b_0 + 1/2(y^T y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \end{aligned} \quad (13)$$

Naturally, removing observations does lead to similar updates for the sufficient statistics:

$$\begin{aligned} \Lambda_0 &= (\Lambda_n - X^T X) \\ \mu_0 &= \Lambda_0^{-1}(\Lambda_n \mu_n - X^T y) \\ a_0 &= a_n - n/2 \\ b_0 &= b_n - 1/2(y^T y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \end{aligned} \quad (14)$$

Later on we will use the term "downdate" to refer to this adjustment by removing observations.

3.2.2 Posterior Predictive

The posterior predictive of the Normal-Inverse-Gamma (NIG) describes the probability of y^* given all previous observations Y which can be summarized directly through the sufficient statistics:

$$p(y^* \mid y) = \int p(y^* \mid \beta, \sigma^2) p(\beta, \sigma^2 \mid y) d\beta d\sigma^2$$
$$= Student_{2a_n}(X\mu_n, \frac{b_n}{a_n}(I + X\Lambda_n^{-1}X^T)) \quad (15)$$

The Student-t distribution is of the multivariate type:

$$Student_\nu(\beta, \mu, \Sigma) = \frac{\Gamma(\frac{1}{2}(\nu + d))}{\Gamma(\nu/2)\pi^{d/2}|\nu\Sigma|^{1/2}}$$
$$\left[1 + \frac{1}{\nu}(\beta - \mu)^T \Sigma^{-1}(\beta - \mu)\right]^{-\frac{1}{2}(\nu+d)} \quad (16)$$

Here d is the dimension of β and μ.

And for completeness sake, the Student-t distribution amounts to:

$$\log p = \log\Gamma(\frac{1}{2}(\nu + d)) - \log\Gamma(\nu/2)$$
$$- \frac{d}{2}\log(\pi) - \frac{1}{2}\log\det(\nu\Sigma)$$
$$- \frac{1}{2}(\nu + d)\log[1 + \frac{1}{\nu}(\beta - \mu)^T \Sigma^{-1}(\beta - \mu)] \quad (17)$$

Note, that in our case Σ is not a matrix, but a scalar.

Also observe that we consistently, collect the independent and dependent variables (x_i, y_i) of a single observation by one random variable w_i.

3.2.3 Sample from the NIG Distribution

To sample from a Normal-Inverse-Gamma distribution, we sample the standard deviation using the Gamma distribution with a and b as hyperparameters:

$$\tau \sim \mathcal{G}(a, b) \quad (18)$$

Then $\sigma = \tau^{-1/2}$. The line coefficients are sampled from a Normal distribution:

$$\mu \sim \mathcal{N}(\mu_0, \sigma^2\Lambda^{-1}) \quad (19)$$

3.3 Extension to Multiple Lines

The extension of Bayesian linear regression can be visualized (Fig. 1) through plate notation (Buntine, 1994). There is Bayesian line regression in parallel for k lines (with k in theory up to infinity). The likelihood function of the full model:

$$p(\pi \mid \alpha) \prod_k \prod_i p(z_i \mid \pi) p(w_i \mid \theta_k) p(\theta_k \mid \lambda_0) \quad (20)$$

In the plate model it can be seen that the cluster proportions π are not integrated out. The Dirichlet process generates a partition π. The partition consists out of indices $z_0, \ldots, z_N$ that link the observations $w_0, \ldots, w_N$ with the parameters $\theta_0, \ldots, \theta_K$. The probability $p(w_i \mid \theta_k)$ corresponds to the likelihood equations 8 and 9 with w_i the tuple of x_i and y_i and θ_k the line parameters σ_k^2 and β_k. The probability $p(\theta_k \mid \lambda_0)$ corresponds to the prior from equation 10. The parameters θ_k (that is, σ_k^2 and β_k) are generated from hyperparameters λ_0. The hyperparameters $\lambda_0 = \{\mu_0, \Lambda_0, a, b\}$ are the parameters from the Normal-Inverse-Gamma prior.

3.4 Gibbs Sampling Parameters

We now consider the Gibbs sampling of the parameters, by which we mean, the sampling of all parameters tied to the observations (not just the unique ones tied to each cluster). The individual steps are described in detail in Algorithm 1. This Gibbs algorithm is known as Algorithm 1 (Neal, 2000).

We perform a loop in which for T iterations each θ_i belonging to observation w_i is updated in sequence. First, the likelihood L_i for all θ_{-i} given w_i is calculated. Second, the posterior predictive for w_i given the hyperparameters $p(w_i \mid \phi_0)$ is calculated. The fraction with the Dirichlet process concentration parameter α subsequently defines if θ_i will be sampled from a new cluster or if one of the existing clusters will be sampled. If a new cluster is sampled, the sufficient statistics are updated with information on w_i and thereafter θ is sampled from a Normal-Inverse-Gamma distribution with the updated hyperparameters.

3.5 Gibbs Sampling Clusters

Directly sampling over the clusters is known as Algorithm 2 (Neal, 2000).

Rather than updating each θ_i per observation w_i, an entire cluster θ_k is updated. In Algorithm 1 the update of a cluster would require a first observation to generate a new cluster at θ_j and then moving all observations of the old cluster θ_i to θ_j.

Algorithm 2 follows the same procedure in excluding w_i from calculating the likelihood. This requires the previously mentioned "downdate" from the corresponding sufficient statistics. In Algorithm 2 after all observations have been iterated over and assigned the corresponding cluster k, an outer loop iterates over all clusters to obtain new parameters θ from the NIG prior.

Algorithm 2: Gibbs sampling over clusters c_k.

procedure GIBBS ALGORITHM 2(w, λ_0, α) ▷ Accepts points w and hyperparameters λ_0 and α, returns k line coordinates
  for all $t = 1 : T$ **do**
    for all $i = 1 : N$ **do**
      c = cluster(w_i) ▷ Get cluster c currently assigned to observation w_i
      λ_c = downdate(w_i, λ_c) ▷ Adjust sufficient statistics for cluster c by removing observation w_i
      $m_c = m_c - 1$ ▷ Adjust cluster size m_c (observation w_i removed reduces it with one)
      for all $k = 1 : K$ **do**
        $L_k = m_k$ likelihood(w_i, θ_k) ▷ Update likelihood for cluster k given observation w_i
      end for
      P_i = post_pred(w_i, λ_0) ▷ Posterior predictive of w_i given hyper parameters
      $p(new) = \frac{\alpha P_i}{\alpha P_i + \sum_k L_k}$ ▷ Sample new or old?
      if $p(new)$ **then**
        λ_k = update(w_i, λ_0) ▷ Update sufficient statistics with observation w_i
        $\theta_i \sim NIG(\lambda)$ ▷ Sample θ_i from NIG
      else
        k sampled from existing clusters
        λ_k = update(w_i, λ_k) ▷ Restore sufficient statistics with observation w_i
      end if
      $m_k = m_k + 1$ ▷ Increment cluster size m_k
    end for
    for all $k = 1 : K$ **do**
      $\theta_k \sim NIG(\lambda_k)$ ▷ Sample θ_k from NIG
    end for
  end for
  return summary on θ_k for k lines
end procedure

4 RESULTS

The Infinite Line Mixture Model (see section 3) is able to fit an infinite number of lines through a point cloud in two dimensions. These lines are no line segments, but infinite lines. However, to test the model a variable number of lines are generated of a length that is considerably larger compared to the spread caused by the standard deviation of points from that line.

As described before, Gibbs sampling leads to correlated samples. We choose to get the Maximum A Posterior estimates for our clusters by picking the median values for all the parameters involved.

4.1 Clustering Performance

The results are measured using conventional metrics for clustering performance. For example the Rand Index describes the accuracy of cluster assignments (Rand, 1971):

$$R = \frac{a+b}{a+b+c+d} \tag{21}$$

Here a numbers the pair of points that belong to the same cluster, both at ground truth as well as after the inference procedure. Likewise b numbers the pair of points that belong to different clusters in both sets. The values c and d describe discrepancies between the ground truth and the results after inference. A Rand Index of one means that there have been no mistakes.

The clustering performance is separate from the line estimation performance. If the points are not properly assigned, the line will not be estimated correctly. Due to the fact that line estimation has this secondary effect, this performance is not taken into account. Moreover, from lines that generated only a single, or very few points, we can extract point assignments, but line coefficients are impossible to derive. This would lead to introducing a threshold for the number of points per cluster. Moreover, the performance would then need to be measured by weighting the fitting versus the assignment.

The performance of Algorithm 1 can be seen in Fig. 2 and is rather disappointing. On average the inference procedure agrees upon the ground truth for 75% of the cases considering the Rand Index. More-

over, if we adjust for chance as with the Adjusted Rand Index, the performance drops to only having 25% correct!

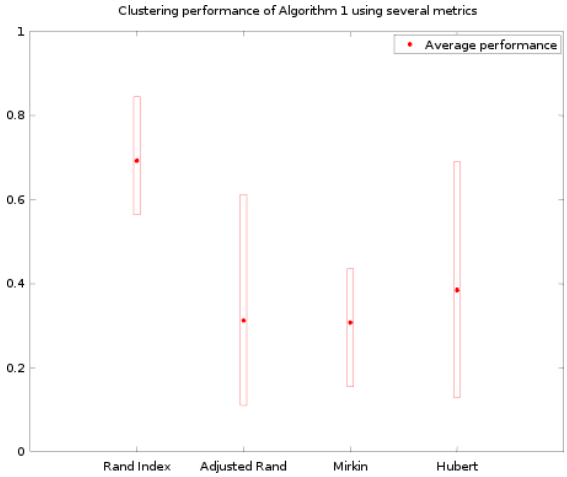

Figure 2: The performance of Algorithm 1 with respect to clustering is measured using the Rand Index, the Adjusted Rand Index, the Mirvin metric, and the Hubert metric. A figure of 1 means perfect clustering for all metrics, except Mirvin's where 0 denotes perfect clustering.

Algorithm 2 leads to stellar performance measures (Fig. 3). Apparently updating entire clusters at once with respect to their parameter values leads at times to perfect clustering, bringing the performance metrics close to their optimal values.

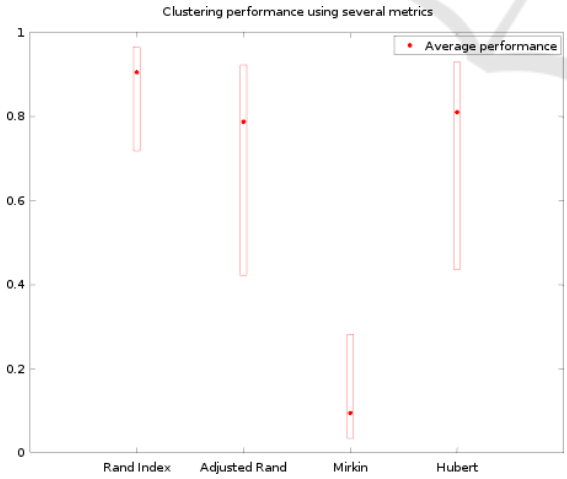

Figure 3: The performance of Algorithm 2 with respect to clustering is measured using the Rand Index, the Adjusted Rand Index, the Mirvin metric, and the Hubert metric. A figure of 1 means perfect clustering for all metrics, except Mirvin's where 0 denotes perfect clustering.

The lack of performance of Algorithm 1 is not only caused by slower mixing (time required to reach the steady state distribution). Also when allowing it ten times the number of iterations of Algorithm 2, it still does not reach the same performance levels. A line seems to form local regions of high probability making it difficult for points to postulate slightly changed line coordinates.

4.2 Some Examples

In the following we show a few examples to understand the inference process better. Figure 4 shows the assignment after a single Gibbs step in Algorithm 1. There is a single line that is represented by two clusters. Algorithm 1 does not have merge or split steps to group these clusters at once, it thus has to move each data point one by one. By the way, there are split-merge algorithms that take these more sophisticated Gibbs steps into account (Jain and Neal, 2004).

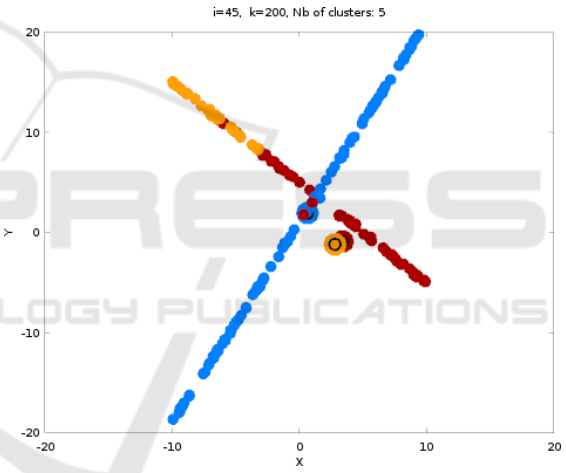

Figure 4: One of the Gibbs steps in the inference of two particular lines. The points are more or less distributed according to the lines, but one line exists out of two large clusters. The line coordinates are visualized by a double circle. The x-coordinate is the y-intercept of the line, the y-coordinate is the slope.

The example in Fig. 5 shows that a single point as an outlier is not a problem for our method. A single point might throw off Bayesian linear regression, but because there are multiple lines to be estimated in our Infinite Line Mixture Model, this single point is assigned its own line.

The extension to more points as outliers would of course require us to postulate a distribution for these outlier points as well. A uniform distribution might for example be used in tandem with the proposed model. This however would lead to a non-conjugate model and hence different inference methods.

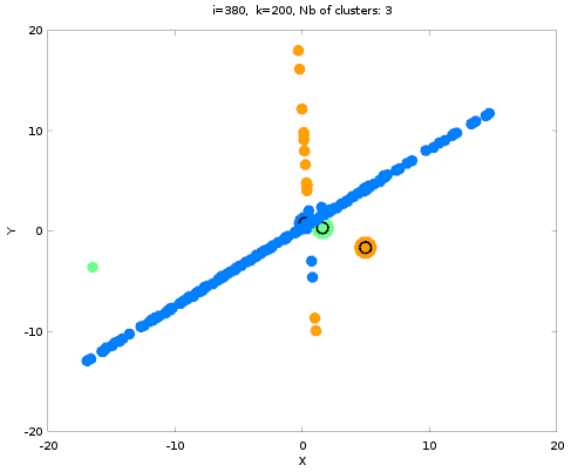

Figure 5: The assignment of a line to a single point. There are three clusters found, rather than only the obvious two.

5 CONCLUSIONS

The Infinite Line Mixture Model that is proposed extends the familiar Bayesian linear regression model to an infinite number of lines using a Dirichlet Process as prior. The model is a full Bayesian method to detect multiple lines. A full Bayesian method, in contrast to ad-hoc methods such as the Hough transform or RANSAC, means optimal inference (Zellner, 1988) given the model and noise definition.

Results in section 4 show high values for difference performance metrics for clustering, such as the Rand Index, the Adjusted Rand Index, and other metrics. The Bayesian model is solved through two types of algorithms. Algorithm 1 iterates over all observations and suffers from slow mixing. The individual updates makes it hard to reassign large number of points at the same time. Algorithm 2 iterates over entire clusters. This allows updates for groups of points leading to much faster mixing. Note, that even optimal inference results in occasional misclassifications. The dataset is generated by a random process. Hence, occassionally two lines are generated with almost the same slope and intercept. Points on these lines are impossible to assign to the proper line.

The essential contribution of this paper is the introduction of a fully Bayesian method to infer lines and there are two ways in which the postulated model can to be extended for full-fledged inference in computer vision as required in robotics. First, the extension of lines in 2D to planes in 3D. This is quite a trivial extension that does not change anything of the model except for the dimension of the data points. Second, somehow a prior needs to be incorporated to limit the lines of infinite length, to line segments. To restrict points on the lines to a uniform distribution of points over a line segment, a symmetric Pareto distribution can be used as prior (for the end points). This would subsequently allow for a hierarchical model in which these end points are in their turn part of more complicated objects. Hence, the Infinite Line Mixture Model is an essential step towards the use of Bayesian methods (and thus properly formulated priors) for robotic computer vision.

REFERENCES

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Bolles, R. C. and Fischler, M. A. (1981). A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In *IJCAI*, volume 1981, pages 637–643.

Bonci, A., Leo, T., and Longhi, S. (2005). A bayesian approach to the hough transform for line detection. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 35(6):945–955.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

Buntine, W. L. (1994). Operations for learning with graphical models. *JAIR*, 2:159–225.

Chen, H., Meer, P., and Tyler, D. E. (2001). Robust regression for data with multiple structures. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–1069. IEEE.

Dahyot, R. (2009). Statistical hough transform. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(8):1502–1509.

de Finetti, B. (1992). Foresight: Its logical laws, its subjective sources. In *Breakthroughs in statistics*, pages 134–174. Springer.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Fienberg, S. E. et al. (2006). When did Bayesian inference become "Bayesian"? *Bayesian analysis*, 1(1):1–40.

Gael, J. V., Teh, Y. W., and Ghahramani, Z. (2009). The infinite factorial hidden markov model. In *Advances in Neural Information Processing Systems*, pages 1697–1704.

Gallo, O., Manduchi, R., and Rafii, A. (2011). CC-RANSAC: Fitting planes in the presence of multiple surfaces in range data. *Pattern Recognition Letters*, 32(3):403–410.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Ghahramani, Z. and Griffiths, T. L. (2005). Infinite latent feature models and the indian buffet process. In

Advances in neural information processing systems, pages 475–482.

Hough, P. V. (1962). Method and means for recognizing complex patterns. Technical report.

Jain, S. and Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1).

Kwon, S.-W., Bosche, F., Kim, C., Haas, C. T., and Liapi, K. A. (2004). Fitting range data to primitives for rapid local 3D modeling using sparse range point clouds. *Automation in Construction*, 13(1):67–81.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Minka, T. (2000). Bayesian linear regression. Technical report, Citeseer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560.

Sudderth, E. B. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*, pages 1585–1592.

Vasudevan, S., Gächter, S., Nguyen, V., and Siegwart, R. (2007). Cognitive maps for mobile robots - an object based approach. *Robotics and Autonomous Systems*, 55(5):359–371.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280.

Zhang, W. and Ksecká, J. (2007). Nonparametric estimation of multiple structures with outliers. In *Dynamical Vision*, pages 60–74. Springer.