

000
001
002
003
004
005
006
007
008
009
010
011054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

Small-Variance Nonparametric Clustering on the Hypersphere

Anonymous CVPR submission

Paper ID 440

Abstract

Structural regularities in man-made environments reflect in the distribution of their surface normals. Describing these surface normal distributions is important in many computer vision applications, such as scene understanding, plane segmentation, and regularization of 3D reconstructions. Based on the small-variance limit of Bayesian nonparametric von-Mises-Fisher mixture distributions, we propose two new flexible and efficient clustering algorithms for directional data, such as surface normals. The first, DP-vMF-means, is a batch clustering algorithm derived from the Dirichlet process vMF mixture. Recognizing the sequential nature of data collection in many applications, we extend this algorithm to the DDP-vMF-means, which infers temporally evolving cluster structure from streaming data. Both algorithms naturally respect the geometry of directional data, which lies on the unit sphere. We demonstrate their performance on synthetic directional data and real 3D surface normal data from RGB-D sensors. While our experiments revolve around 3D data, both algorithms generalize to higher dimensional directional data such as protein backbone configurations and semantic word vectors.

1. Introduction

Man-made environments and objects exhibit clear structural regularities such as planar or rounded surfaces. These properties are evident on all scales from small objects such as books over tables, rooms and buildings to the organization of whole cities. Such regularities can be captured in the statistics of surface normals that describe the local differential structure of a shape. These statistics contain valuable information that can be used for scene understanding, plane segmentation, or to regularize a 3D reconstruction.

Inference algorithms in fields such as robotics or augmented reality, which would benefit from the use of surface normal statistics, are not generally provided the whole batch of data a priori. Instead, they are often provided a stream of data batches from structured light or time-of-flight depth cameras. Thus, capturing the surface normal statistics of

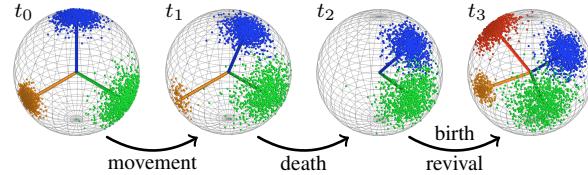


Figure 1: Evolution of the data distribution on the sphere which we model using a Dependent Dirichlet Process von-Mises-Fisher mixture model (DDP-vMF-MM).

man-made structures often necessitates the temporal integration of observations from a vast data stream of varying cluster mixtures. Additionally, such applications pose hard constraints on the amount of computational power available, as well as tight timing constraints.

We address these challenges by focusing on flexible Bayesian nonparametric (BNP) Dirichlet process mixture models (DP-MM) which describe the distribution of surface normals in their natural space, the unit sphere in 3D. Taking the small variance asymptotic limit of this DP-MM of von-Mises-Fisher (vMF) distributions, we obtain a fast k -means-like algorithm, which we call DP-vMF-means, to perform non-parametric clustering of data on the unit hypersphere. Furthermore, we propose a novel dependent Dirichlet process of von-Mises-Fisher distributions to achieve integration of directional data into a temporally consistent streaming model. Recognizing the need for efficient inference algorithms, we additionally derive k -means-like algorithms via the small variance asymptotic limit of these distributions. Finally, we propose a method, inspired by optimistic concurrency control, for parallelizing the inherently sequential labeling process of BNP-derived algorithms. This allows the real-time operation of the proposed algorithms.

Beyond the aforementioned vision applications, directional data is ubiquitous in many other fields including protein backbone configurations in computational biology, semantic word vectors in natural language processing, and rotations expressed as quaternions in robotics. Further, many of those sources are observed as a stream. The proposed algorithms directly generalize to these other data sources as well as higher-dimensional spherical data.

108

2. Related Work

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

where \mathcal{I}_k is the set of data indices assigned to cluster k , and where the conjugate prior for μ_k yields $p(x_i; \mu_0, \tau_0, \tau)$ via marginalization over μ_k ,

$$\begin{aligned} p(x_i; \mu_0, \tau_0) &= \int_{\mu_k} \text{vMF}(x_i; \mu_k, \tau) \text{vMF}(\mu_k; \mu_0, \tau_0) d\mu_k \\ &= \frac{Z(\tau)Z(\tau_0)}{Z(\|\tau x_i + \tau_0 \mu_0\|_2)}. \end{aligned} \quad (5)$$

4.1. DP-vMF-means

Label Update: To derive a hyperspherical analog to DP-means [17], consider the limit of the label sampling step (4) as $\tau \rightarrow \infty$. The normalizer $Z(\|\tau x_i + \tau_0 \mu_0\|_2)$ approaches

$$Z(\|\tau x_i + \tau_0 \mu_0\|_2) \xrightarrow{\tau \rightarrow \infty} \exp(-\tau) \quad (6)$$

where overscript $\tau^{(*)}$ denotes proportionality up to a finite power of τ , and where we have used the fact that as $\tau \rightarrow \infty$, the modified Bessel function of the first kind satisfies [1]

$$I_{D/2-1}(\tau) = \frac{\exp(\tau)}{\sqrt{2\pi\tau}} \left(1 - O\left(\frac{1}{\tau}\right)\right) \xrightarrow{\tau \rightarrow \infty} \exp(\tau). \quad (7)$$

To achieve a nontrivial result, the asymptotic behavior of $Z(\tau)$ must be matched by α , so let $\alpha = \exp(\lambda\tau)$ to obtain

$$\frac{\alpha Z(\tau)Z(\tau_0)}{Z(\|\tau x_i + \tau_0 \mu_0\|_2)} \xrightarrow{\tau \rightarrow \infty} Z(\tau) \exp(\tau(\lambda + 1)). \quad (8)$$

Therefore, as $\tau \rightarrow \infty$, the label sampling step becomes

$$\begin{aligned} \lim_{\tau \rightarrow \infty} p(z_i = k | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x}) &= \lim_{\tau \rightarrow \infty} \begin{cases} \frac{|\mathcal{I}_k| e^{\tau(x_i^T \mu_k - \lambda - 1)}}{\sum_{j=1}^K |\mathcal{I}_j| e^{\tau(x_i^T \mu_j - \lambda - 1)} + c(\tau)} & k \leq K \\ \frac{c(\tau)}{\sum_{j=1}^K |\mathcal{I}_j| e^{\tau(x_i^T \mu_j - \lambda - 1)} + c(\tau)} & k = K + 1. \end{cases} \end{aligned} \quad (9)$$

where we have used that the normalizers $Z(\tau)$ of $\text{vMF}(x_i; \mu_k, \tau)$ and Eq. (8) cancel and $c(\tau) \xrightarrow{\tau \rightarrow \infty} 1$. Thus, as $\tau \rightarrow \infty$, sampling from $p(z_i | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x})$ is equivalent to the following assignment rule:

$$z_i = \arg \max_{k \in \{1, \dots, K+1\}} \begin{cases} x_i^T \mu_k & k \leq K \\ \lambda + 1 & k = K + 1. \end{cases} \quad (10)$$

Note that since $-1 \leq x_i^T \mu_k \leq 1$, the parameter λ can be restricted to the set $\lambda \in [-2, 0]$ without loss of generality. Doing so elucidates the intuitive meaning of λ ; the parameter is equivalent to a maximum angular spread ϕ_λ of clusters about their mean direction, via $\lambda = \cos(\phi_\lambda) - 1$.

Parameter Update: Taking the limit as $\tau \rightarrow \infty$ of the parameter posterior for cluster k from Eq. (2), the hyperparameters τ_0 and μ_0 become negligible. Hence the parameter update becomes:

$$\mu_k = \frac{\sum_{i \in \mathcal{I}_k} x_i}{\|\sum_{i \in \mathcal{I}_k} x_i\|_2} \forall k \in \{1, \dots, K\}. \quad (11)$$

Cost Function: From Eq. (10) we can see that assigning a datapoint x_i to cluster k provides a score of $x_i^T \mu_k$, whereas adding a new cluster provides a score of $\lambda + 1 - x_i^T \mu_{K+1} = \lambda$, since new mean directions are initialized directly to $\mu_{K+1} = x_i$. Hence, the cost function that the DP-vMF-means algorithm optimizes is

$$J_{\text{DP-vMF}} = \max \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} x_i^T \mu_k + \lambda K. \quad (12)$$

Algorithmically, when a datapoint is assigned to a new cluster, i.e. $z_i = K + 1$, we initialize the mean of that cluster to $\mu_{K+1} = x_i$. When updating labels, if an observation is the last one in its cluster, the cluster is removed prior to finding the new label for that observation using Eq. (10).

5. Dependent Dirichlet Process vMF-MM

Suppose now that, in addition to an unknown number K of components, the vMF mixture undergoes temporal evolution in discrete timesteps $t = 1, 2, \dots$ consisting of three stochastic processes: Mixture components can move, be destroyed, and new ones can be created at each timestep. For such a scenario, the dependent Dirichlet process (DDP) [18] is an appropriate prior over the mixture components and weights. It constructs a Markov chain of Dirichlet processes G_t , where G_{t+1} is sampled from G_t as follows:

1. **(Death)** For each atom θ in G_t , sample from Bernoulli(q). If the result is 1, add θ to F_0 .
2. **(Motion)** Replace each θ in F_0 with $\theta' \sim T(\theta' | \theta)$.
3. **(Birth)** Sample a DP $F_1 \sim \text{DP}(\alpha, H)$. Let G_{t+1} be a random convex combination of F_0 and F_1 .

There are four parameters in this model: $\alpha > 0$ and $H(\cdot)$, the concentration parameter and base measure of the innovation process; $q \in (0, 1)$, the Bernoulli cluster survival probability; and finally $T(\cdot | \cdot)$, the random walk transition distribution. In the present work, both the base and random transition distributions are von-Mises-Fisher: $H(\mu) = \text{vMF}(\mu; \mu_0, \tau_0)$, and $T(\mu | \nu) = \text{vMF}(\mu; \nu, \xi)$.

Suppose at timestep t , a new batch of data \mathbf{x} is observed. Then Gibbs sampling posterior inference for the DDP mixture, as in the previous sections, iteratively samples labels and parameters. Let the set of tracked mean directions from previous timesteps be $\{\mu_{k0}\}_{k=1}^K$, where $\mu_{k0} \sim \text{vMF}(m_k, \tau_k)$ and Δt_k denotes the number of timesteps since cluster k was last instantiated (i.e. when it last had

324 data assigned to it).¹ Then the label sampling distribution is
 325

$$326 p(z_i = k | \mu, \mathbf{z}_{-i}, \mathbf{x}) \\ 327 \propto \begin{cases} \alpha^{\frac{1-q^t}{1-q}} p(x_i; \mu_0, \tau_0) & k = K + 1 \\ (c_k + |\mathcal{I}_k|) \text{vMF}(x_i; \mu_k, \tau) & k \leq K, |\mathcal{I}_k| > 0 \\ q^{\Delta t_k} c_k p(x_i; m_k, \tau_k) & k \leq K, |\mathcal{I}_k| = 0, \end{cases} \quad (13)$$

331 where c_k is the number of observations assigned to cluster
 332 k in past timesteps. The parameter sampling distribution is
 333

$$334 p(\mu_k | \mu_{-k}, \mathbf{z}, \mathbf{x}) \propto \begin{cases} \text{vMF}(\mu_k; \frac{\mu'_k}{\|\mu'_k\|_2}, \|\mu'_k\|_2) & c_k = 0 \\ p(\mu_k | \mathbf{x}, \mathbf{z}; m_k, \tau_k) & c_k > 0, \end{cases} \quad (14)$$

337 where $\mu'_k = \tau_0 \mu_0 + \tau \sum_{i \in \mathcal{I}_k} x_i$, and $p(\mu_k | \mathbf{x}, \mathbf{z}; m_k, \tau_k)$ is
 338 the distribution over the current cluster k mean direction μ_k given the assigned datapoints and information about the old
 339 mean direction μ_{k0} .
 340

341 5.1. DDP-vMF-means

343 In the following, we analyze the small-variance asymptotics
 344 of the DDP-vMF mixture model. We first derive the
 345 label assignment rules, followed by the parameter updates.
 346

347 **Label Update:** First, let $\alpha = \exp \lambda \tau$, $q = \exp Q \tau$, $\xi =$
 348 $\exp \beta \tau$, and $\tau_k = \tau w_k$, with $\lambda \in [-2, 0]$ as before, $Q \leq 0$,
 349 and $\beta, w_k \geq 0$. Note that $\lim_{\tau \rightarrow \infty} \frac{1-q^t}{1-q} = 1$, and thus the
 350 asymptotics of the label assignment probability for current
 351 and new clusters is the same as in Section 4.1.

352 Hence, we focus on the assignment of a datapoint to a
 353 previously observed, but currently not instantiated, cluster
 354 k . During the Δt_k timesteps since cluster k was last ob-
 355 served, the mean direction μ_k underwent a random vMF
 356 walk $\mu_{k0} \rightarrow \mu_{k1} \rightarrow \dots \rightarrow \mu_{k\Delta t_k} = \mu_k$ with initial dis-
 357 tribution $\mu_{k0} \sim \text{vMF}(\mu_{k0}; m_k, \tau_k)$. Therefore, the interme-
 358 diate mean directions $\{\mu_{kn}\}_{n=1}^{\Delta t_k}$ must be marginalized out
 359 when computing $p(x_i; m_k, \tau_k)$:

$$361 p(x_i; m_k, \tau_k) \\ 362 = \int \dots \int_{\mu_{k0}, \dots, \mu_{k\Delta t_k}} p(x_i | \mu_{k\Delta t_k}; \tau) \cdot p(\mu_0; m_k, \tau_k) \\ 363 \cdot \prod_{n=1}^{\Delta t_k} p(\mu_{kn} | \mu_{k(n-1)}; \xi) \\ 364 = Z(\tau) Z(\beta \tau)^{\Delta t_k} Z(w_k \tau) \int \dots \int_{\mu_{k0}, \dots, \mu_{k\Delta t_k}} \exp(\tau f) \quad (15)$$

$$367 f = x_i^T \mu_{k\Delta t_k} + \beta \sum_{n=1}^{\Delta t_k} \mu_{kn}^T \mu_{k(n-1)} + w_k \mu_{k0}^T m_k.$$

370 The integration in (15) cannot be computed in closed
 371 form; however, the value of the integral is only of interest
 372 in the limit as $\tau \rightarrow \infty$. Therefore, Theorem 1, an extension
 373 of Laplace’s approximation to general differentiable mani-
 374 folds, may be used to obtain an exact formula.

375 ¹Note that all quantities ($\mathbf{x}, \mu_{k0}, m_k, \tau_k, c_k, n_k$, etc.) are now time-
 376 varying. This dependence is not shown in the notation for brevity, and all
 377 quantities are assumed to be shown for the current timestep t .

378 **Theorem 1** (Manifold Laplace Approximation). Suppose
 379 $M \subset \mathbb{R}^n$ is a bounded m -dimensional differentiable mani-
 380 fold and $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth function on M . Fur-
 381 ther, suppose f has a unique global maximum on M , $x^* =$
 382 $\arg \max_{x \in M} f(x)$. Then
 383

$$384 \lim_{\tau \rightarrow \infty} \frac{\int_M e^{\tau f(x)} dM}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla^2 f(x^*) U|^{\frac{1}{2}} e^{\tau f(x^*)}} = 1 \quad (16)$$

385 where $U \in \mathbb{R}^{n \times m}$ is a matrix whose columns are an or-
 386 thonormal basis for the tangent space of M at x^* .
 387

388 *Proof.* See the supplementary material. The general tech-
 389 nique of this proof is to transform coordinates between
 390 the manifold and its tangent plane using the exponential
 391 map [8], and then apply the multidimensional Laplace ap-
 392 proximation in the transformed Euclidean space. \square
 393

394 **Corollary 1.** Given a smooth function $f: (\mathbb{R}^D)^N \rightarrow \mathbb{R}$, with
 395 a unique global maximum over the N -product of $(D - 1)$ -
 396 spheres $x^* \in (\mathbb{S}^{D-1})^N$,

$$397 \int_{(\mathbb{S}^{D-1})^N} e^{\tau f(x)} \xrightarrow{\tau \rightarrow \infty} e^{\tau f(x^*)}. \quad (17)$$

400 *Proof.* The result is a special case of Theorem 1 on a par-
 401 ticular manifold. \square
 402

403 Using Corollary 1 and the limiting approximation of the
 404 modified Bessel function (7) in equation (15) yields the fol-
 405 lowing asymptotic behavior as $\tau \rightarrow \infty$:

$$406 p(x_i; m_k, \tau_k) \xrightarrow{\tau \rightarrow \infty} \exp(\tau(f^* - 1 - \beta \Delta t_k - w_k)). \quad (18)$$

407 The only remaining unknown in the asymptotic expres-
 408 sion, f^* , can be found via constrained optimization
 409

$$410 \max_{\{\mu_{kn}\}_{n=1}^{\Delta t_k}} x_i^T \mu_{k\Delta t_k} + \beta \sum_{n=1}^{\Delta t_k} \mu_{kn}^T \mu_{k(n-1)} + w_k \mu_{k0}^T m_k \quad (19)$$

$$411 \text{s.t. } \mu_{kn}^T \mu_{kn} = 1 \forall n \in \{0, \dots, \Delta t_k\}.$$

412 The optimization (19) has a closed-form solution:
 413

$$414 \mu_{k0} = \frac{w_k m_k + \beta \mu_{k1}}{\|w_k m_k + \beta \mu_{k1}\|_2} \\ 415 \mu_{kn} = \frac{\mu_{k(n+1)} + \mu_{k(n-1)}}{\|\mu_{k(n+1)} + \mu_{k(n-1)}\|_2} \forall n \in \{1, \dots, \Delta t_k - 1\} \quad (20) \\ 416 \mu_{k\Delta t_k} = \frac{x_i + \beta \mu_{k(\Delta t_k - 1)}}{\|x_i + \beta \mu_{k(\Delta t_k - 1)}\|_2}.$$

417 These ternary relationships enforce that the optimal vMF
 418 mean directions along the random walk lie on the geodesic
 419 between m_k and x_i . Therefore, this walk can be described
 420

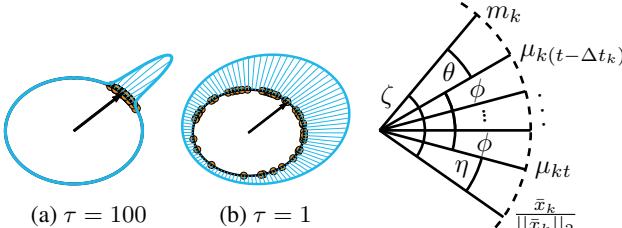


Figure 2: Left: 2D vMF distributions. Right: geometry of the state transition distribution maximum likelihood setting of $\mu_{k0}, \mu_{k1}, \dots, \mu_{k\Delta t_k}$. When assigning a label for data-point i , note that $\bar{x}_k/\|\bar{x}_k\|_2 = x_i$.

geometrically by three angles, as shown in Fig. 2: the angle ϕ between consecutive μ_{kn} , the angle η between x_i and $\mu_{k\Delta t_k}$, and the angle θ between m_k and μ_{k0} . Given these definitions, standard trigonometry yields a set of three equations in ϕ , η , and θ :

$$\begin{aligned} w_k \sin(\theta^*) &= \beta \sin(\phi^*) = \sin(\eta^*) \\ \zeta &= \theta^* + \Delta t_k \phi^* + \eta^* = \arccos(m_k^T x_i) \end{aligned} \quad (21)$$

where ζ is defined as the full angle between x_i and m_k . Since (21) cannot be solved in closed-form, Newton's method is used to compute ϕ^* , θ^* , and η^* , which in turn determines f^* :

$$f^* = w_k \cos(\theta^*) + \beta \Delta t_k \cos(\phi^*) + \cos(\eta^*). \quad (22)$$

Returning to (18), the asymptotics of the transition distribution are

$$p(x_i; m_k; \tau_k) \xrightarrow{\tau \rightarrow \infty} \exp \left(\frac{\tau w_k (\cos(\theta^*) - 1)}{+\tau \beta \Delta t_k (\cos(\phi^*) - 1)} \right). \quad (23)$$

Substituting this into Eq. (13) with the earlier definition $q = \exp(\tau Q)$, and taking the limit $\tau \rightarrow \infty$ yields the assignment rule $z_i = \arg \max_k J_k$ where

$$J_k = \begin{cases} \lambda + 1 & k = K + 1 \\ \mu_k^T x_i & k \leq K, |\mathcal{I}_k| > 0 \\ \left(\begin{array}{l} \Delta t_k \beta (\cos(\phi^*) - 1) \\ + w_k (\cos(\theta^*) - 1) \\ + \cos(\eta^*) + \Delta t_k Q \end{array} \right) & k \leq K, |\mathcal{I}_k| = 0. \end{cases} \quad (24)$$

Note that if $Q \Delta t_k < \lambda$, then the old cluster can be removed permanently as it will never be revived. Furthermore, the parameter Q can be constrained to $Q \in [\lambda, 0]$ without loss of generality, since the behavior of the algorithm is the same for all $Q \leq \lambda$; all clusters are immediately removed after each timestep, and the algorithm reduces to DP-vMF-means.

Parameter Update: The parameter update rule for DDP-vMF-means comes from the asymptotic behavior of (14) as $\tau \rightarrow \infty$. The analysis for any new cluster is the same as that in Section 4.1, so our focus is again on the transitioned mean direction posterior $p(\mu_{k\Delta t_k} | \mathbf{x}, \mathbf{z}; m_k, \tau_k)$ (recall that $\mu_k = \mu_{k\Delta t_k}$ in the definition of the random vMF walk). This distribution can be expanded, similarly to (15), as:

$$\begin{aligned} p(\mu_{k\Delta t_k} | \mathbf{x}, \mathbf{z}; m_k, \tau_k) &= \int \cdots \int_{\mu_{k0}, \dots, \mu_{k(\Delta t_k-1)}} p(\mathbf{x} | \mu_{k\Delta t_k}; \tau) \cdot p(\mu_0; m_k, \tau_k) \\ &\quad \cdot \prod_{n=1}^{\Delta t_k} p(\mu_{kn} | \mu_{k(n-1)}; \xi) \\ &= Z(\tau)^{|\mathcal{I}_k|} Z(\beta \tau)^{\Delta t_k} Z(w_k \tau) \int \cdots \int_{\mu_{k0}, \dots, \mu_{k(\Delta t_k-1)}} \exp(\tau f) \end{aligned} \quad (25)$$

$$f = \sum_{i \in \mathcal{I}_k} x_i^T \mu_{k\Delta t_k} + \beta \sum_{n=1}^{\Delta t_k} \mu_{kn}^T \mu_{k(n-1)} + w_k \mu_{k0}^T m_k.$$

Define $\bar{x}_k = \sum_{i \in \mathcal{I}_k} x_i$. Once again, applying Corollary 1 in the limit $\tau \rightarrow \infty$ removes the integrals over the marginalized mean directions. However, in contrast to the label assignment update, $\mu_{k\Delta t_k}$ is not marginalized out; therefore, an additional maximization with respect to $\mu_{k\Delta t_k}$ to find the concentration point of the posterior as $\tau \rightarrow \infty$ yields

$$\begin{aligned} p(\mu_{k\Delta t_k} | \mathbf{x}, \mathbf{z}; m_k, \tau_k) &\xrightarrow{\tau \rightarrow \infty} \\ &\exp(\tau(f^* - |\mathcal{I}_k| - \Delta t_k \beta - w_k)) \\ f^* &= w_k \cos(\theta^*) + \beta \Delta t_k \cos(\phi^*) + \|\bar{x}_k\|_2 \cos(\eta^*). \end{aligned} \quad (26)$$

Again, analyzing the geometry of the geodesic between $\bar{x}_k/\|\bar{x}_k\|_2$ and m_k (see Fig. 2) there exist ϕ^* , θ^* and η^* such that

$$\begin{aligned} w_k \sin(\theta^*) &= \beta \sin(\phi^*) = \|\bar{x}_k\|_2 \sin(\eta^*) \\ \zeta &= \theta^* + \Delta t_k \phi^* + \eta^* = \arccos(m_k^T \frac{\bar{x}_k}{\|\bar{x}_k\|_2}), \end{aligned} \quad (27)$$

which can be solved via Newton's method. Given the solution, μ_k can be obtained by rotating $\frac{\bar{x}_k}{\|\bar{x}_k\|_2}$ by angle η^* on the geodesic shown in Fig. 2 towards m_k ,

$$\mu_k = R(\eta^*) \frac{\bar{x}_k}{\|\bar{x}_k\|_2}. \quad (28)$$

Weight Update: After the iteration of label and parameter updates has converged, the weight w_k must be updated for all clusters to reflect the new uncertainty in the mean direction of cluster k . This can be done by examining (26): Since at the maximum of a vMF($\mu; m_k, w_k \tau$) density, $\exp(\tau w_k m_k^T \mu) = \exp(\tau w_k)$, w_k is updated to f^* .

6. Optimistic Iterated Restarts (OIR)

In our implementation of the algorithm we pay special attention to speed and parallel execution to enable real-time performance for streaming RGB-D data.

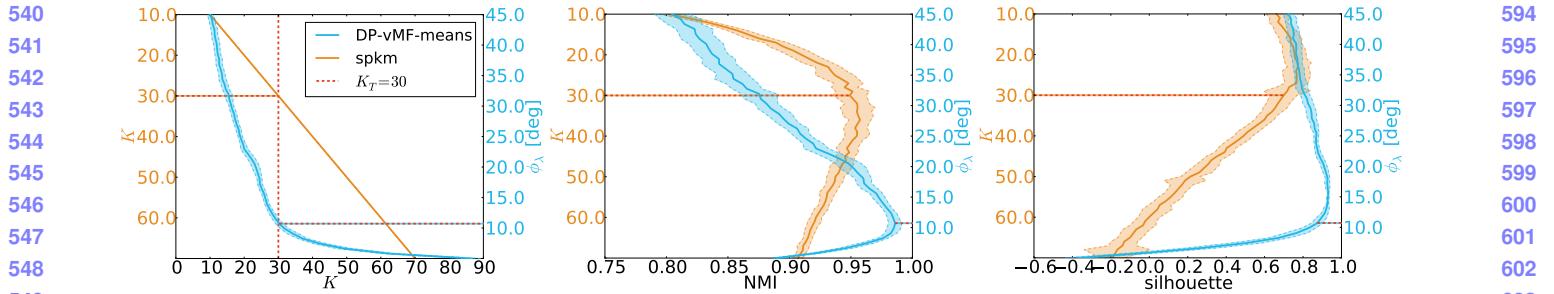


Figure 3: Comparison of the parametric spkm and the nonparametric DP-vMF-means clustering algorithms on synthetic spherical data with $K_T = 30$ clusters. Note DP-vMF-means’ higher maximum normalized Mutual Information (NMI) as well as silhouette score. Red dotted lines indicate the parameters achieving the true number of clusters.

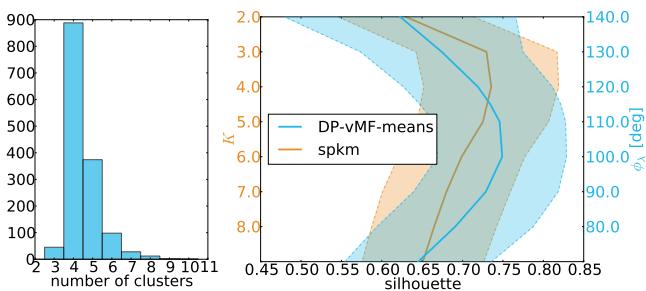


Figure 4: Left: histogram over the number of clusters found by the DP-vMF-means algorithm for all scenes in the NYU v2 RGB-D dataset [20]. Right: average silhouette values for DP-vMF-means as well as spkm.

Observe that the main bottle-neck in the DP-based hard clustering algorithms is the assignment of labels. This is because it is an inherently sequential process, as the label assignment for each observation depends on all previous assignments. We address this issue with an optimistic parallel label assignment procedure inspired by techniques for database concurrency control [21]. Note that OIR generalizes to other BNP-based hard clustering algorithms such as the DP-means [17] or the Dynamic means [6] algorithms.

First, we compute assignments in parallel (e.g. on a GPU). If datapoints were assigned only to instantiated clusters, we output the labeling. Otherwise, we find the lowest observation id i that modified the number of clusters, and recompute the assignments for all observations $i' > i$ in parallel. We repeat this until no new clusters are created and no old ones are revived.

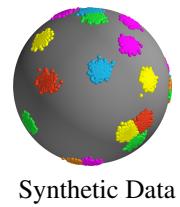
Note that the DP-vMF-means algorithm requires K restarts for a batch of data, where K is the final number of clusters inferred by the algorithm. For the DDP-vMF-means algorithm we expect K restarts over the whole stream of data, where K is the sum of total number of clusters created and the total number of revivals of old clusters. In practice, as displayed in Fig. 6, we observe that there are few restarts, since the DDP-vMF-means tracks clusters consistently over time and rarely adds or revives clusters.

7. Results

7.1. Evaluation of the DP-vMF-means Algorithm

Synthetic Data: First, we evaluate the behavior of the DP-vMF-means algorithm in comparison to its parametric cousin, the spkm algorithm, on synthetic 3D spherical data sampled from $K_T = 30$ true vMF distributions. All evaluation results are shown as the mean and standard deviation over 50 runs. The left plot of Fig. 3 depicts the inferred number of clusters K on the horizontal axis as a function of the respective parameters of the two algorithms: the number of clusters K for spkm and the parameter ϕ_λ for DP-vMF-means (recall that $\phi_\lambda = \cos^{-1}(\lambda + 1)$ as defined in Sec. 4.1). This figure demonstrates the ability of the DP-vMF-means algorithm to discover the correct number of clusters K_T , and the relative insensitivity of the discovered number of clusters with respect to its parameter ϕ_λ .

The middle and right hand plots show two measures for clustering quality. The normalized Mutual Information (NMI) [27], depicted in the middle, is computed using the true labels. DP-vMF-means achieves an almost perfect NMI of 0.99, while spkm only reaches 0.96 NMI even with $K = K_T$. The slightly superior performance of DP-vMF-means stems from its enhanced ability to avoid local optima due to the way labels are initialized: while spkm is forced to initialize K cluster parameters, DP-vMF-means starts with an empty set and adds clusters on the fly as more data are labelled. The NMI results are corroborated by the silhouette score [24], shown to the right in Fig. 3. Ranging from -1 to 1 , the silhouette score is an internal measure for clustering quality that can be computed without knowledge of the true clustering, and is widely used to tune parametric clustering algorithms such as k -means. With a maximum of 0.93 DP-vMF-means reaches a close to perfect silhouette score, indicating well-separated, concentrated clusters. Again, spkm does not reach the same clustering performance even for $K = K_T$ for the same aforementioned reasons.



Synthetic Data



Figure 5: Directional segmentation of scenes from the NYU v2 RGB-D dataset [20] as implied by surface normal clusters. The complexity of the scenes increases from left to right as can be observed from the RGB images in the top row. The second row shows the clustering inferred using DP-vMF-Means with $\phi_\lambda = 100^\circ$ while the third and fourth show the spherical k -means results for comparison. Black denotes missing data due to sensor limitations. Note that DP-vMF-means adapts the number of clusters to the complexity of the scene.

NYU v2 depth dataset: In this experiment, the DP-vMF-means and spkm algorithms were compared on the NYU v2 RGB-D dataset [20]. Surface normals were extracted directly from the depth images [14] and preprocessed with total variation smoothing [23]. We quantify the clustering quality in terms of the average silhouette score over the clusterings of the 1449 scenes of the NYU v2 depth dataset. Since we do not possess the true scene labeling, we use the silhouette quality metric as a proxy for the NMI metric; this was motivated by the agreement between the maxima of the NMI and silhouette scores in the previous synthetic experiment.

Across the whole NYU v2 dataset, the DP-vMF-means algorithm achieves the highest average silhouette score of 0.75 for $\phi_\lambda^* = 100^\circ$ as depicted in Fig. 4. The histogram over the number of inferred clusters by DP-vMF-means for ϕ_λ^* indicates the varying complexity of the scenes ranging from three to eleven. The clear peak at $K = 4$ coincides with the highest silhouette score for spkm. This explains the only slightly lower silhouette score of spkm since most scenes in the dataset seem to exhibit four clusters in their surface normal distribution.

Figure 5 shows a qualitative comparison of the scene segmentation implied by the clustering of surface normals. In comparison to spkm, the DP-vMF-means clustering results show the ability of the algorithm to adapt the number of clusters to the scene at hand. If the right number of clusters is selected for the spkm clustering, the results have sim-

ilar quality; however, the number of clusters is generally not known a priori, and thus the parametrization in terms of K is the major drawback of the spkm algorithm in comparison to DP-vMF-means.

7.2. Evaluation of the DDP-vMF-means Algorithm

Real-time Directional Segmentation: In fields such as mobile robotics or augmented reality, it is uncommon to observe just a single RGB-D frame of a scene; more typically, the sensor will observe a temporal sequence of frames. The proposed DDP-vMF-means algorithm is designed to handle such streams of directional data in a temporally consistent manner. The following experiment is a demonstration of a key ability of the DDP-vMF-means algorithm: the ability to *revive* clusters, thereby maintaining a consistent labeling despite periods of not observing all of the clusters. We compare against the ad-hoc approaches of clustering on a frame-by-frame basis using DP-vMF-means, both with and without initializing the algorithm from the previous frame's clusters (the former is referred to as sequential DP-vMF-means). Motivated by the DP-vMF-means evaluation, all three algorithms were run with $\phi_\lambda = 100^\circ$. For DDP-vMF-means $\beta = 10^5$ and $Q = \frac{\lambda}{400}$ were used.

Before computing surface normals from each depth image, the image was preprocessed using edge-preserving smoothing with a guided filter [13]. Choosing the guidance image equivalent to the input image results in a $O(N)$ filter that provides similar smoothing results to a bilateral filter,

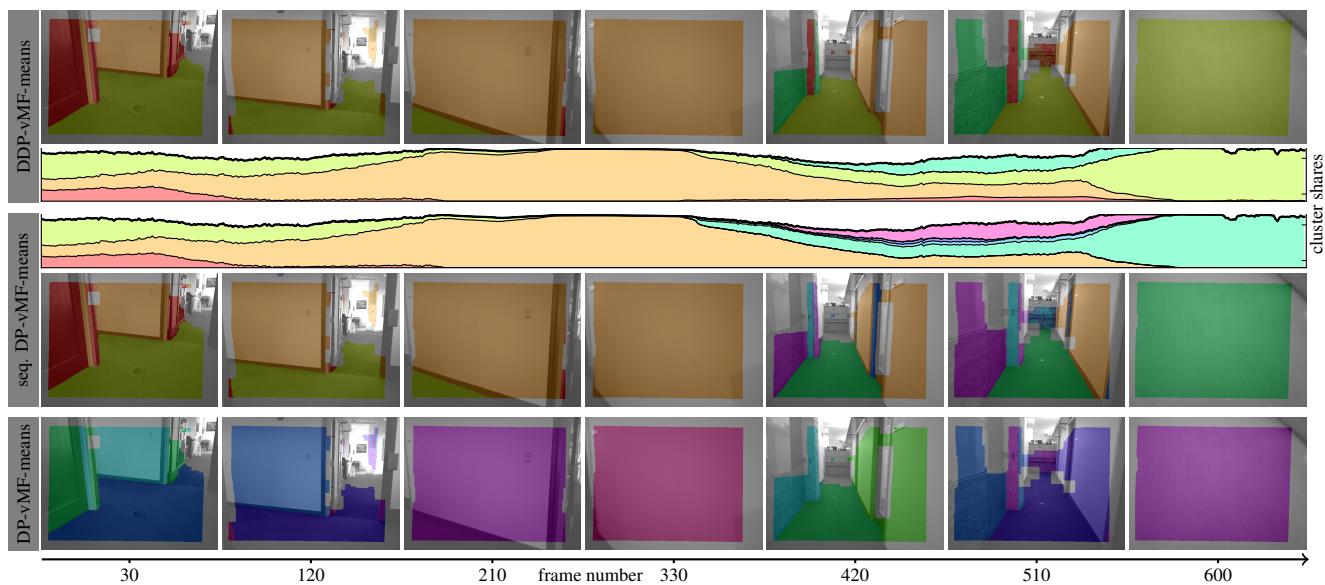


Figure 6: Clustering of a surface normal stream recorded when walking a 90° turn in an office environment. We depict key-frames color-coded with the implied surface-normal clustering for three different clustering algorithms. The plots in the second and third row depict the percentage of normals associated to the respective cluster for the DDP-vMF-means as well as sequential DP-vMF-means. Note that the clustering obtained via the DDP-vMF-means algorithm is consistent across the whole run as opposed to the other algorithms.

which is $O(Nr^2)$, where N is the number of pixels and r the filter radius. The resulting hybrid CPU-GPU implementation of the guided filter processes a 640×480 depth image in 10 ms using a filter window of 20×20 pixels.

Three variants of algorithms are considered: DP-vMF-means purely frame-by-frame, DP-vMF-means with initialization from previous frame and DDP-vMF-means. Initializing DP-vMF-means from the previous frame's cluster centers allows a greedy frame-to-frame label consistency. However, unlike DDP-vMF-means, reinstating previous clusters with multiframe lapses is not possible. This can be observed in Fig. 6, which shows the percentage of normals associated with a specific cluster. While DDP-vMF-means is temporally consistent and reinstates the lime-green and red clusters, observed in the first half of the run, DP-vMF-means erroneously creates new clusters. We do not depict the percentages of surface normals associated with the clusters for the batch DP-vMF-means algorithm, since there is no label consistency between time-steps as can be observed in the last row of Fig. 6.

The average run-time per frame was 27.0 ms for batch DP-vMF-means, 13.5 ms for sequential DP-vMF-means, and 14.0 ms for DDP-vMF-means. The increased running time of Batch DP-vMF-means is a result of clustering each batch of surface normals in isolation; OIR label assignment needs several restarts to assign labels to all surface normals. By initializing the clusters from a previous frame, the sequential DP-vMF-means only incurs labeling restarts if a

new cluster is observed, and hence has significantly lower run time. DDP-vMF-means is slightly slower than sequential DP-vMF-means since it is keeping track of both observed and unobserved clusters.

8. Conclusion

Taking the small-variance asymptotic limit of the Bayesian nonparametric DP-vMF as well as the DDP-vMF mixture model, we have derived two novel spherical-kmeans-like algorithms for efficient batch and streaming clustering on the unit hypersphere.

The DP-vMF-means demonstrates its flexibility when clustering the surface normal distribution of the whole NYU v2 RGB-D dataset. Interestingly, the highest clustering quality according to the silhouette score is attained for a maximal cluster size of angle $\phi_\lambda = 100^\circ$. This reflects the typical 90° angles between structural elements in man-made environments, plus 10° slack for sensor noise.

Implementing Optimistic Iterated Restarts (OIR) parallelizes label assignments, leading to real-time performance for DDP-vMF-means clustering on batches of $\sim 300,000$ surface normals collected at 30 Hz from a depth camera.

We envision a large number of potential applications for the presented algorithms in computer vision and beyond. Their implementation will be made available at [URL withheld for double blind review].

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

References

- [1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover Books on Mathematics. Dover Publications, 1965. 2, 3
- [2] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway. Clustering on the unit hypersphere using von Mises-Fisher distributions. *JMLR*, 6(9), 2005. 2
- [3] M. Bangert, P. Hennig, and U. Oelfke. Using an infinite von Mises-Fisher mixture model to cluster treatment beam directions in external radiation therapy. In *ICMLA*, pages 746–751. IEEE, 2010. 2
- [4] C. Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974. 2
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003. 2
- [6] T. Campbell, M. Liu, B. Kulis, J. P. How, and L. Carin. Dynamic clustering via asymptotics of the dependent Dirichlet process mixture. In *NIPS*, pages 449–457, 2013. 2, 6
- [7] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175, 2001. 2
- [8] M. P. do Carmo. *Riemannian Geometry*. Birkhäuser Verlag, Boston, MA, 1992. 4
- [9] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1995. 2
- [10] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422–1429. IEEE, 2009. 2
- [11] S. Gopal and Y. Yang. von Mises-Fisher clustering models. In *ICML*, pages 154–162, 2014. 2
- [12] M. A. Hasnat, O. Alata, and A. Trémeau. Hierarchical 3-d von Mises-Fisher mixture model. In *Workshop on Divergences and Divergence Learning, ICML*, 2013. 2
- [13] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, pages 1–14. Springer, 2010. 7
- [14] D. Holz, S. Holzer, and R. B. Rusu. Real-Time Plane Segmentation using RGB-D Cameras. In *Proceedings of the RoboCup Symposium*, 2011. 7
- [15] K. Jiang, B. Kulis, and M. Jordan. Small-variance asymptotics for exponential family Dirichlet process mixture models. In *NIPS*, 2012. 2
- [16] J. T. Kent. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society*, pages 71–80, 1982. 2
- [17] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *ICML*, 2012. 2, 3, 6
- [18] D. Lin, E. Grimson, and J. Fisher. Construction of dependent Dirichlet processes based on poisson processes. *NIPS*, 2010. 3
- [19] K. V. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009. 2
- [20] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 6, 7
- [21] X. Pan, J. E. Gonzalez, S. Jegelka, T. Broderick, and M. Jordan. Optimistic concurrency control for distributed unsupervised learning. In *NIPS*, pages 1403–1411, 2013. 6
- [22] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney. Spherical topic models. In *ICML*, pages 903–910, 2010. 2
- [23] G. Rosman, Y. Wang, X.-C. Tai, R. Kimmel, and A. M. Bruckstein. Fast regularization of matrix-valued images. In *ECCV*, volume 7574, pages 173–186. Springer, 2012. 7
- [24] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 6
- [25] A. Roychowdhury, K. Jiang, and B. Kulis. Small-variance asymptotics for hidden Markov models. In *NIPS*, 2013. 2
- [26] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher III. A mixture of manhattan frames: Beyond the manhattan world. In *CVPR*, 2014. 2
- [27] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *JMLR*, 3:583–617, 2003. 6
- [28] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, New York, 2010. 2
- [29] R. Triebel, W. Burgard, and F. Dellaert. Using hierarchical em to extract planes from 3d range scans. In *ICRA*, pages 4437–4442. IEEE, 2005. 2
- [30] S. Zhong. Efficient online spherical k-means clustering. In *IJCNN*, volume 5, pages 3180–3185. IEEE, 2005. 2
- 918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971