修士論文

混合分布による道路状態推定

(Estimation of Road Traffic Condition using Mixture Distribution Model)

山本　敬介

東京大学大学院 情報理工学系研究科 電子情報学専攻

指導教員 安達 淳

2013 年 2 月 6 日提出

概要

人々が扱うデータは日々巨大化している．また，そのデータを解析することで得られる有用な情報によって，我々の生活も劇的に変化してきている．ただ，こういった有用な情報を得るためには，解析を行うためにデータを保存しておく必要がある．中でも，常時データが流入してくるストリームデータに関しては，そのデータ量に上限はなく，無限に流入してくるため保存することが難しい．外部記憶装置に保存するには，大量の装置を用意するための豊富な資本力が必要となり，現実的ではない．また，限られた外部記憶装置に保存する方法として，時系列に沿って必要な情報のみを残し，それ以外の情報を破棄する研究も行われている．しかしそれでは，必要度の高いデータは得られても，データ全体を把握することはできない．そこで本論文では，ストリームデータに対して，データの圧縮を実現しつつデータ全体の再現を行うことができる手法を提案する．

本研究では，対象とするデータをストリームデータであるプローブカーデータとし，その平常時の道路状態を混合正規分布モデルを用いて表す．プローブカーデータとは，車載センサーから得られる速度や位置の情報であり，現在搭載車は増加しつつある．そして推定されたモデルを用いて平常状態のデータを破棄し，圧縮を行うことを目標とする．道路状態は，道路の種別や時間帯によって変化する．また，同じ特徴量であっても，ある箇所ではそれが平常データである一方で，他の箇所ではそれが異常データであることがある．そこで，位置情報と時間情報を用いてプローブカーデータをセグメントに分割する．これにより，セグメントに適応したモデルを推定する．モデルの推定には，EM アルゴリズムを用いる．混合正規分布の平均，分散，混合比率を推定するが，平均と分散に関しては，全てのセグメントで共通のものを用いることにする．共通混合正規分布にすることによって，個々の正規分布が道路状態を表すと捉えることができ，セグメントごとに混合比率を推定することで，そのセグメントがどのような道路状態なのかということがわかる．こうして推定したセグメントごとの個別混合正規分布は，平常状態を表すモデルとなっている．このモデルを用いて，プローブカーデータの圧縮を行う．圧縮を行うにあたって，単にデータ量を削減するだけではなく，元デー

タが再現可能となるような圧縮を行う．個別混合正規分布モデルから再現が可能となるようなデータを破棄することによって，データを圧縮する．事前にある閾値を定めておき，データが流入した際に，流入データの特徴量とその特徴量における個別混合正規分布モデルとを比べ，モデルの確率密度が閾値より大きければ，流入データを破棄する．また，平常状態モデルから外れているような異常データについては圧縮を行わず保持する．こうすることにより，後の解析において有益となる異常データを残しつつ，元データの復元も可能な圧縮を行う．

目次

図目次

表目次

第 1 章
序論

1.1 研究の背景

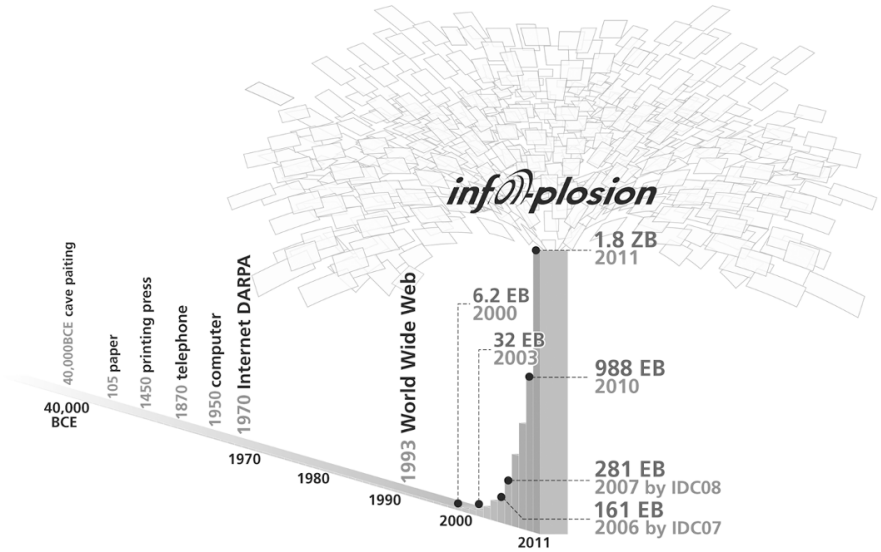

図 1.1: 情報爆発時代　出処:特定領域研究『情報爆発時代に向けた新しい IT 基盤技術の研究』

情報技術の発展と共に，流通，蓄積されるデータ量は日々巨大化している．図.1.1からもわかるように，年々その増大の勢いを増している．中でも，通常のデータベースで扱うことが難しい程の巨大なデータをビッグデータと呼び，近年その活用が叫ばれている．ビッグデータの具体例として気象や GPS 等のセンサーデータ，オンラインショッピングの購買履歴，インターネット上に保存された文章やデジタル写真などが挙げられる．これらのデータを分析することで，天候や人や車両の流れ，売り上げの増加，トレンドの特定などの大きなメリットがあり，それらは我々の生活を大きく変えつつある．しかし，データの膨大さから，利用するデータを蓄積するのには限界がある．とりわけ，ストリームデータにおいては全てを蓄積するのは現実的ではない．ストリームデータというのは，Cyber Physical System(CPS) データに代表される，常時情報が入ってくるデータである．CPS とは，車や人，気象など実世界をデジタルデータとして取り込み，活用するシステムのことである．加速度センサーのついた携帯電話や，車載 GPS，気象情報センサーなどが身近なものとしてある．こうした CPS データは位置情報や，時刻，温度などであり，これらのデータを解析することで，混雑状況や，リアルタイムの気象情報などの有益な情報を得るわけである．それらの過去のデータを解析し，現在に活かそうとする場合，過去のデータを蓄積する必要があるが，ストリームデータの場合，全

てを蓄積することは不可能である．例えば，100 万個のノードから毎秒 100byte のデータが送られてくるストリームデータを考える．1 秒当たりでは 100Mbyte だが，1 日では約 8.6Tbyte にもなり，1 年分のデータを蓄積すると，約 3.1Pbyte となり，全てを蓄積するのは現実的に不可能となる．そのためデータを破棄，もしくは圧縮する必要が出てくるが，ストリームデータには，平常時のデータにはあまり有用性はなく，時折発生する異常を検知するなど，値が異常であるものが有用であることがある．例えば気象情報センサーであれば，普段は晴れや雨などの天候と判断できる平常時の値であるが，台風やフェーン現象が発生した時などは風速や，気温が異常な値を示す．こういった異常値が得られた場合に，それがどういう現象であるのかという判断することが重要である．また，特定の現象が起こった場合に得られた値を蓄積しておくことが，後に分析することを考えると有用である．このようなデータの場合，両者を区別せず破棄することは好ましくない．このようなデータに対し，元データの情報を損なうことなくデータを圧縮する方法が求められている．

1.2 本研究の目的

そこで我々は，全てを蓄積することが非現実的であるストリームデータに対し，平常時のデータを学習し事前にモデルを作り，平常時のデータを圧縮することで効率的なデータの蓄積を行った．今回，ストリームデータであるプローブカーデータに着目しデータの圧縮を行った．全国を小さな区画に区切り，そこに所属する車の特徴を学習し混合分布モデル化することで，区画ごとの道路状態を推定する．こうして平常状態をモデル化することで，大半のデータはこのモデルから得られると考えることができ，データを蓄積する必要がなくなる．新たにデータが取得された場合，該当する区画のモデルに照らし合わせ，もっともらしいデータであれば破棄，そうでないもののみを蓄積する．こうすることにより，有用なデータは蓄積しつつデータの圧縮を行った．

1.3 本論文の構成

本論文は，全 6 章から構成される．第 2 章では，本研究で用いる Kullback-Leibler Divergence(KL-Divergence)，EM アルゴリズムによるパラメタ推定，モデルの考え方の元となる Latent Dirichlet Allocation(LDA) について説明する．第 3 章では，提

案手法である EM アルゴリズムを用いたモデル推定と，モデルを用いた圧縮法を説明する．第 4 章では，今回用いるプローブカーデータについて説明し，第 5 章ではプローブカーデータを用いて実験を行い，提案手法の有効性を示す．最後に第 6 章で本論文をまとめ，考察を行う．

第 2 章
関連研究

2.1 はじめに

本章では，本研究で用いる関連手法である Kullback-Leibler Divergence，EM アルゴリズムを用いた混合ガウス分布のパラメータ推定，Latent Dirichlet Allocation(LDA) について紹介する．

2.2 Kullback-Leibler Divergence

ここでは，データを生成する真の確率分布を近似するモデルの良さを評価する基準として用いる Kullback-Leibler Divergence(KL-Divergence) とその性質について述べる．

未知の確率分布関数 $G(x)$ に従って観測された n 個のデータを $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ とする．データを発生するこの確率分布関数 $G(x)$ を以下では真の分布と呼ぶことにする．これに対して推定されたモデルを $F(x)$ とする．確率分布関数 $G(x)$ 及び $F(x)$ が，それぞれ密度関数 $g(x)$ 及び $f(x)$ をもつ場合は連続分布モデルという．一方，$g(x)$ 及び $f(x)$ が有限もしくは加算無限個の離散点 $\{x_1, x_2, \cdots, x_k, \cdots\}$ に対して，次のように事象 $\{\omega; X(\omega) = x_i\}$ の確率

$$
\begin{aligned}
g_i &= g(x_i) \equiv \Pr(\{\omega; X(\omega) = x_i\}) \\
f_i &= f(x_i) \equiv \Pr(\{\omega; X(\omega) = x_i\}) \quad i = 1, 2, \cdots
\end{aligned}
$$

で表される場合は，離散分布モデルという．このとき，モデル $f(x)$ のよさを，真のモデル $g(x)$ との確率分布としての近さによって評価するものとする．この近さを測る尺度として，KL-Divergence は以下のように定義される．

$$
I(G; F) = E_G\left[\log \frac{G(X)}{F(X)}\right] \tag{2.1}
$$

ここで，E_G は確率分布 G に関する期待値を示す．確率分布関数が密度関数 $g(x)$ と $f(x)$ をもつ連続モデルの場合には，KL-Divergence は

$$
I(g; f) = \int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx \tag{2.2}
$$

と表される．一方，確率が $\{g(x_i); i = 1, 2, \cdots\}, \{f(x_i); i = 1, 2, \cdots\}$ で与えられる離

散モデルの場合には,

$$I(g;f) = \sum_{i=1}^{\infty} g(x_i) \log \frac{g(x_i)}{f(x_i)} \tag{2.3}$$

と表される.

KL-Divergence には，次のような性質がある.

(i) $I(g;f) \geq 0$

(ii) $I(g;f) = 0 \Leftrightarrow g(x) = f(x)$

この性質から，KL-Divergence の値が小さいほど，モデル $f(x)$ は $g(x)$ に近いと考えることができる.

2.3 混合ガウス分布のパラメタ推定

潜在変数を持つモデルの最尤解を求めるための 1 つの方法として，EM アルゴリズムがある．混合ガウスモデルが与えられている時，各要素の平均と分散，そして混合係数からなるパラメタについて尤度関数を最大化するような値を求める.

まず，尤度関数を求める．k 番目のガウシアンの平均を $\boldsymbol{\mu}_k$，分散を $\boldsymbol{\Sigma}_k$，混合係数を θ_k とすると，混合ガウス分布が K 個のガウス分布の次の形の線形重ね合わせで書ける.

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \theta_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.4}$$

ここで,

$$\sum_{k=1}^{K} \theta_k = 1 \tag{2.5}$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{2\pi\Sigma_k}} \exp -\frac{(x-\mu_k)^2}{2\Sigma_k} \tag{2.6}$$

である．この混合ガウス分布を，観測したデータ集合 $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ にあてはめる問題を考える．このデータ集合は，第 n 行を $\boldsymbol{v}_n^T$ とする $N \times D$ 行列 $\boldsymbol{V}$ と表される．対数尤度関数は以下のように表される.

$$\ln P(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \theta_k \mathcal{N}(\boldsymbol{v}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{2.7}$$

この式について，平均 $\boldsymbol{\mu}_k$，分散 $\boldsymbol{\Sigma}_k$ そして混合係数 θ_k を初期化し，対数尤度の初期値を計算しする．ガウス要素の平均 $\boldsymbol{\mu}_k$ に関して微分して 0 とおくと，次式が得られる．

$$0 = \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{2.8}$$

ここで $\gamma(z_{nk})$ は事後確率と呼ばれ，以下の式で表される．

$$\gamma(z_{nk}) = \frac{\theta_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \theta_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{2.9}$$

式 (2.8) を整理すると，

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{2.10}$$

となる．式 (2.7) の $\boldsymbol{\Sigma}_k$ に関する微分を 0 とおき，整理すると次式が得られる．

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{2.11}$$

ここで $\boldsymbol{\mu}_k$ は，式 (2.10) で推定されたものを用いる．最後に $\ln P(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$ を $\boldsymbol{\theta}_k$ について最大化する．ここに，各要素の総和が 1 であるという制約条件 (2.6) を考慮しなくてはならない．よってラグランジュ未定係数法を用いる．すなわち，次の量を $\theta_k (k = 1, \ldots, K)$ で微分して 0 とおく．

$$\ln p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \theta_k - 1 \right) \tag{2.12}$$

すると次式が得られる．

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \theta_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{2.13}$$

ここで再び事後確率が出現している．両辺に θ_k を掛けて k について和を取り，制約条件 (2.6) を用いると $\lambda = -N$ を得る．これを用いて λ を消去して変形すると，

次式を得る．

$$\theta_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \tag{2.14}$$

すなわち，k 番目の要素に関する混合係数は，その要素の全データ点に対する事後確率の平均で与えられる．EM アルゴリズムにおいては，現在のパラメタ値を用いて事後確率を計算することを Expectation ステップ (E ステップ)，その事後確率を用いてパラメタ値を再計算することを Maximization(M ステップ) と呼ぶ．推定される値が収束するまで，この E ステップ，M ステップを繰り返し，値を求める．M ステップの後，対数尤度関数の変化またはパラメタの変化量がある閾値より小さくなったときにアルゴリズムが収束したと判断し更新を終了する．また，一般に対数尤度には多くの極大解が存在し，EM アルゴリズムはそれらのうちで最大のものに収束するとは限らない．そのため，初期値の設定にばらつきを持たせ，何度か実行することにより最も対数尤度の大きい値を求めることになる．

2.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation(LDA) は文書コーパスのようなデータ集合の確率的生成モデルである．LDA における文書モデルとは，まず語の生成確率分布を与えるモデルとして複数のトピックを考え，そして文書をトピックの混合モデルと考えるというものである．つまり，その文書で表現したい内容によって，どのトピックが好まれるかが決まり，好みに従って選ばれたトピックを表現する単語がランダムに選ばれる．文書モデルはトピックの重ねあわせで表され，その混合比率で特徴付けられる．トピックにおける語の生成確率分布によって，語の意味的な関連性を表現することができる．また文書をトピックの重ねあわせとすることで文書を完結に記述できる．この特徴から，LDA は文書分類や要約などに利用されている．

まず文書 $\boldsymbol{d}$ は複数の語 $\{w_1, \cdots, w_{|d|}\}$ からなるとする．語の生成モデルとして複数のトピックを考え，語はいずれかのトピックから生成されるとする．すなわち，文書中の各単語に対してそれを生成したトピックがあり，観測できる文書中の語の他に，隠れ変数として 1 つ 1 つの語を生成したトピックがある．また，各トピックからの語の生成確率分布はトピックごとに異なる．これにより語の意味的な背景を表現できる．例えば,「音楽」や「絵画」という語の生成確率が高いトピックは,「芸術」というトピックだと考えることができ,「修論」や「教師」といった単

語の生成確率が高いトピックは「教育」というトピックだと考えることができる．例として図.2.1 に,「教育」というトピックにおける語の生成確率分布を示した.「教育」というトピックから生成される語の生成確率が高くなっている「修論」や「教師」という語が多い文書であれば，教育に関する文書であるということになる．しかし，文書は一つのトピックだけでなく，いくつかのトピックを含むことが多い．その場合の例を図.2.2 に示した．この例では,「教育」,「芸術」,「スポーツ」のトピックがあり，各文書がどのような内容であるかを，トピックの重ねあわせで表現している．このように，文書は複数のトピックの重ねあわせであり，その内容によって各トピックの重みを変えることで，その混合比率が文書の内容を表す特徴量となる．例えば，学校の部活動に関する文書であれば,「教育」と「スポーツ」の重みが大きくなるという具合である．

以上から，K 個のトピック $\{t_1, \cdots, t_K\}$，N 個の語彙 $\{v_1, \ldots, v_N\}$ のもとで，M 個の文書からなる文書集合 $\{\boldsymbol{d}_1, \cdots, \boldsymbol{d}_M\}$ の生成は次のようにモデル化される．

- 全ての $\boldsymbol{d}_i$ について
  1. ディリクレ分布 $\mathcal{DIR}(\boldsymbol{\alpha})$ に従って，トピックの分布のパためた $\boldsymbol{\theta}_i$ を生成する
  2. 全ての $w_{ij} \in \boldsymbol{d}_i$ について
     (i) 多項分布 $\mathcal{MULTI}(\boldsymbol{\theta}_i)$ に従ってトピックを選択する．選択されたトピックを t_k とする
     (ii) $z_{ij} \Leftarrow t_k$
     (iii) 多項分布 $\mathcal{MULTI}(\boldsymbol{\phi}_k)$ によって単語を生成する．生成された単語を v_l とする
     (iv) $w_{ij} \Leftarrow v_l$

ここで w_{ij} は i 番目の文書の j 番目の語を表す確率変数であり，z_{ij} は w_{ij} を生成するトピックを表す確率変数である．また $\boldsymbol{\phi}_k$ はトピック t_k が語を生成する確率分布のパラメタであり，$\boldsymbol{\beta}$ をパラメタとするディリクレ分布 $\mathcal{DIR}(\boldsymbol{\beta})$ に従う．

この生成の過程をグラフィカルモデルで表すと図 2.3 のようになる．図からもわかるように，LDA は 3 段階の階層からなるモデルである．$\boldsymbol{\alpha}, \boldsymbol{\beta}$ はコーパスレベルのパラメタであり，最初に決定されその後は普遍である．$\boldsymbol{\theta}_i$ は文書レベルの変数であり，文書ごとに 1 度生成される．z_{ij}, w_{ij} は単語レベルの変数で，各単語ごとに 1 度生成される．

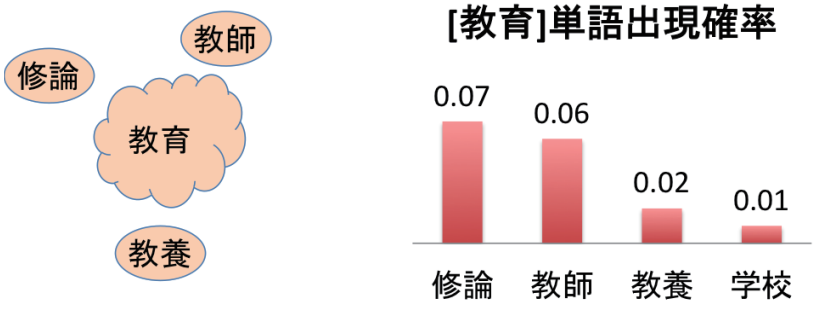

図 2.1: LDA におけるトピックと単語生成多項分布

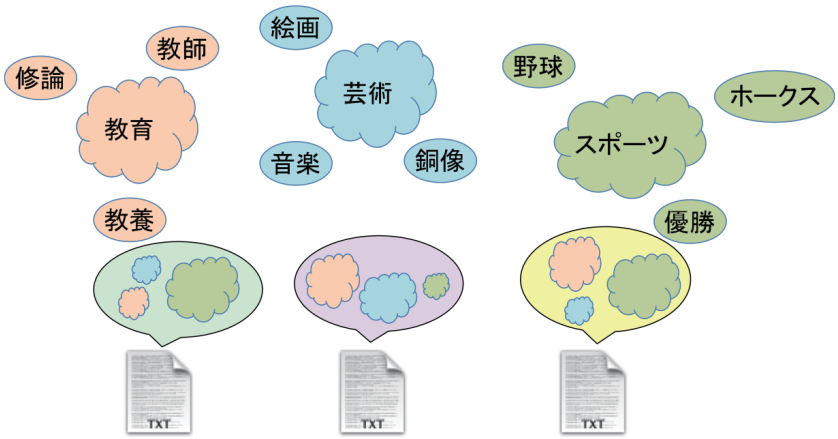

図 2.2: LDA における文書トピック推定

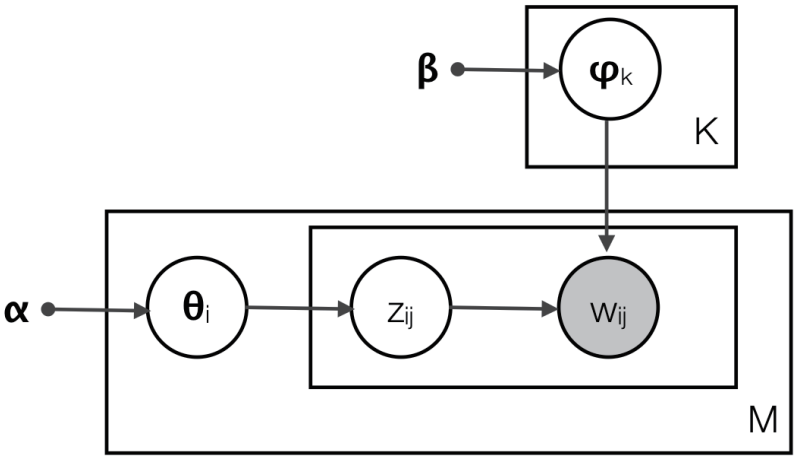

図 2.3: LDA のグラフィカルモデル

第 3 章
パラメタ推定と圧縮

3.1 はじめに

本研究では，蓄積することが非現実的なストリームデータに対し圧縮を行う．セグメント，時間帯で区切られたデータに対し，道路状態モデルをそれぞれ推定し，推定されたモデルを用いて平常データを破棄し圧縮を行う．以下の 3 つのステップにより，これを実現する．

- モデルの定義
- パラメタ推定
- データ圧縮

まず，道路状態モデルにて用いる変数の定義を行う．このとき，データをセグメント，時間帯で分割することで，実際の道路状況に合わせた推定が行われるようにした．次に，平常状態のデータを機械学習し，各セグメントごとの道路状態モデルを推定する．モデルの学習は，EM アルゴリズムを用いて正規混合分布でモデル化した．このとき，正規混合分布の平均，分散は共通のものを用い，混合比率はセグメントごとに推定した．これは，文書ごとに混合比率を変えることで，より精緻なモデルを推定する LDA と同様の考えであり，各セグメントの混合比率を低次元の特徴として扱うことを狙ったものである．本手法と LDA との対比を表 3.1 に示す．LDA では，分布に混合多項分布を用いている．これは対象データに，単語の生成確率という離散データを用いているためで，今回本研究では実数データを対象にするため，混合正規分布を利用する．推定法は，混合正規分布のパラメタ推定として一般的な最尤推定を用いる．LDA では，文書ごとにトピックの混合比率を変えることで，混合比率が文書の内容を低次元で簡潔に表す特徴量となっている．本研究では，これと同様の考えを取り入れ，セグメントごとに混合比率を変えることで，混合比率が道路状態を簡潔に表す特徴量となっているとみなす．

最後に推定されたモデルを用いたデータ圧縮を行う．推定されたモデルにより，平常状態のデータを蓄積する必要をなくし，モデルから外れる異常データとモデルのみを蓄積することでデータ圧縮を行った．データの圧縮では閾値を定め，推定したモデルと照らし合わせて出現率の低いデータのみを蓄積し，それ以外を破棄することでデータを大きく圧縮した．この章では，平常状態のデータを表すモデルを定義し，EM アルゴリズムを用いたパラメタ推定法と，モデルによるデータの圧縮について説明する．

表 3.1: 関連手法と提案手法との対比

	LDA	提案手法
対象データ	離散データ	実数データ
利用する分布	混合多項分布	混合正規分布
推定法	ベイズ推定	最尤推定
特徴	文書ごとにトピックの混合比率を変えた	セグメントごとに, 正規混合分布の混合比率を変える

3.2 モデル

観測されるデータを以下のように定義する

- 道路セグメントの集合 S：道路はあらかじめ適当なサイズのセグメントに分割されているとする
- 時間帯の集合 T：時間は一日を周期とする時間帯にあらかじめ分割されているとする
- 個々のプローブカーデータの特徴ベクトル $\boldsymbol{v} := (v_1, v_2, \cdots, v_d)$：個々の車のデータは，$d$ 次元の特徴ベクトルで表される
- プローブカーデータ $D = \{\boldsymbol{d}_{s_i t_j}\}$：カープローブデータの特徴ベクトル $\boldsymbol{v}$ は，各セグメント $s \in S$ 及び各時間帯 $t \in T$ ごとに，特徴ベクトルの集合 $\boldsymbol{d}_{st}$ にまとめられているものとする

また，各セグメントの状態の集合 $K := \{k_i\}$ を考える．各車の特徴ベクトル $\boldsymbol{d}$ が観測される確率は，セグメントの状態ごとに定義されると仮定する．

$$P(\boldsymbol{v}|k)$$

本研究では，特徴ベクトルの各特徴量は実数値であり，その確率分布は，独立した正規分布に従うと仮定する．

$$P(\boldsymbol{v}|k) = \prod_{i=1}^{d} \mathcal{N}(v_i|\mu_{ki}, \Sigma_{ki}) \tag{3.1}$$

ここで μ_{ki}, Σ_{ki} は，それぞれセグメント状態 k における i 番目の特徴量の平均と分散を表す．また，$N(;\mu, \Sigma)$ は平均 μ，分散 Σ の正規分布を表す．一方，各時間帯 t

における道路セグメント s の道路状況を，道路状態 K の混合状態と考える．混合比率を $\theta_{st} := (\theta_{st1}, \theta_{st2}, \cdots, \theta_{stK})$ で表す．なお，

$$\sum_{k=1}^{K} \theta_{stk} = 1$$

が成り立つ．以上の潜在構造を仮定したとき，道路セグメント s での時間帯 t における特徴ベクトル $\boldsymbol{v}$ を持つ車が観測される確率は，以下の式で表される．

$$\begin{aligned} P(\boldsymbol{v}|t, s) &= \sum_{k \in K} \theta_{st} P(\boldsymbol{v}|k) \\ &= \sum_{k \in K} \theta_{st} \prod_{i=1}^{d} \mathcal{N}(v_i|\mu_{ki}, \Sigma_{ki}) \end{aligned}$$

また，観測データ D の尤度は以下の式で表される．

$$L(D) = \prod_{s \in S} \prod_{t \in T} P(\boldsymbol{d}_{st}) = \prod_{s \in S} \prod_{t \in T} \prod_{v \in d_{st}} P(\boldsymbol{v}|t, s) \tag{3.2}$$

混合ガウス分布のモデルでは，すべての時空間で同一の混合比を用いるが，提案手法では時空間ごとに個別の混合比を用いる．これは，Latent Dirichlet Allocation(LDA) で導入された文書ごとに混合比率を変えることで，より精緻なモデルが得られたのと同様の考えで，各時空間の混合比率を低次元の特徴として使うことを狙ったものである．

3.3 パラメタ推定

ここでは EM アルゴリズムを用いて，前節で述べたモデルのパラメタ推定を行うアルゴリズムを導出する．なお，議論を簡単にするため，各車の特徴ベクトルは一次元とする．各特徴量が独立であるという仮定のもとでは，多次元への拡張は容易である．観測データの集合 D が与えられた時に，上記のモデルのパラメタである

- セグメント s，時間帯 t の道路状態の混合比率 $\{\boldsymbol{\theta}_{st}\}$：ここで $\boldsymbol{\theta}_{st} = (\theta_{st1}, \ldots, \theta_{stK})$ で $\sum_{k=1}^{K} \theta_{stk} = 1$ が成り立つ．
- 各道路状態における各特徴量の正規分布の平均 $\{\mu_k\}$ 及び $\{\Sigma_k\}$

を推定する．これらのパラメタをまとめて $\mathbf{\Lambda}$ で表すことにする．EM アルゴリズムでは，繰り返し計算の中でパラメタを逐次更新していく．以下では $t-1$ 回目の更新で得られるパラメタを $\hat{\mathbf{\Lambda}}$，t 回目で求めるパラメタを $\mathbf{\Lambda}$ と表すことにする．観測データの尤度はパラメタの値によって異なるため，尤度はパラメタを用いて $P(v|t,s;\mathbf{\Lambda}), P(\boldsymbol{d}_{st};\mathbf{\Lambda})$ と表す．

EM アルゴリズムでは，以下の量を考える．

$$Q(\hat{\mathbf{\Lambda}},\mathbf{\Lambda},D) := \sum_{s\in S}\sum_{t\in T}\sum_{v\in\boldsymbol{d}_{st}}\sum_{k=1}^{K} P(k|v,s,t;\hat{\mathbf{\Lambda}}) \cdot \log P(v,k|s,t,\mathbf{\Lambda})$$

この式で事後確率 $P(k|v,s,t;\hat{\mathbf{\Lambda}})$ は，セグメント s，時刻 t においてデータ v が観測された時の状態が k の確率を表しており，以下の式で求めることができる．

$$\begin{aligned} P(k|v,s,t;\hat{\mathbf{\Lambda}}) &= \frac{P(k,v,s,t;\hat{\mathbf{\Lambda}})}{\sum_{k'=1}^{K} P(k',v,s,t;\hat{\mathbf{\Lambda}})} \\ &= \frac{\hat{\theta}_{stk} N(v;\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{k'=1}^{K}\hat{\theta}_{stk'} N(v;\hat{\mu}_{k'},\hat{\Sigma}_{k'})} \end{aligned} \tag{3.3}$$

EM アルゴリズムでは各繰り返しで，上記の量を最大とするパラメタを求める．つまり

$$\arg\max_{\mathbf{\Lambda}} Q(\hat{\mathbf{\Lambda}},\mathbf{\Lambda},D)$$

を解く．各セグメントと各時間帯における混合比率分布については，和が 1 になる必要があるため，ラグランジュ未定乗数法を使って以下の式を考える．

$$\begin{aligned} F(\hat{\mathbf{\Lambda}},\mathbf{\Lambda},D) &:= Q(\hat{\mathbf{\Lambda}},\mathbf{\Lambda},D) + \sum_{s\in S}\sum_{t\in T}\lambda_{st}\left(\sum_{k=1}^{K}\theta_{stk}-1\right) \\ &= \sum_{s,t}\sum_{v\in\boldsymbol{d}_{st}}\sum_{k=1}^{K} P(k|v,s,t;\hat{\mathbf{\Lambda}})\left[\log\theta_{stk} - \frac{\log\Sigma_k}{2}\right. \\ &\quad \left. -\frac{(v-\mu_k)^2}{2\Sigma_k} + const\right] + \sum_{s,t}\lambda_{st}\left(\sum_{k=1}^{K}\theta_{stk}-1\right) \end{aligned} \tag{3.4}$$

各道路状態 k について，正規分布の平均は，以下のようにして求められる．

$$\frac{\partial F(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda}, D)}{\partial \mu_k} = \sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t; \hat{\mathbf{\Lambda}}) \frac{v - \mu_k}{2\Sigma_k} = 0$$

より，μ_k について解くと

$$\mu_k = \frac{\sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t; \hat{\mathbf{\Lambda}}) \cdot v}{\sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t; \hat{\mathbf{\Lambda}})} \tag{3.5}$$

となる．この値は，式 (3.3) で与えられる事後分布を確率分布とすると，特徴量の期待値となっている．同様に，道路状態 k について，正規分布の分散を求める．

$$\frac{\partial F(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda}, D)}{\partial \Sigma_k} = \sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t; \hat{\mathbf{\Lambda}}) \left[-\frac{1}{2\Sigma_k} + \frac{(v - \mu_k)^2}{2\Sigma_k^2} \right]$$

これが 0 に等しいことより，

$$\Sigma_k = \frac{\sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t, ; \hat{\mathbf{\Lambda}}) \cdot (v - \mu_k)^2}{\sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t, ; \hat{\mathbf{\Lambda}})} \tag{3.6}$$

となる．μ_k は式 (3.5) で得られる値を用いる．この値は，式 (3.3) で与えられる事後分布を確率分布とする平均と観測値の差の二乗の期待値となっている．道路セグメント s および時刻 t における混合比率を求める．

$$\frac{\partial F(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda}, D)}{\partial \theta_{stk}} = \sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t; \hat{\mathbf{\Lambda}}) \cdot \frac{1}{\theta_{stk}} + \lambda_{st} = 0$$

より，

$$\theta_{stk} = \frac{\sum_{v \in \boldsymbol{d}_{st}} P(k|v, s, t, ; \hat{\mathbf{\Lambda}})}{\sum_{v \in \boldsymbol{d}_{st}} \sum_{k'=1}^{K} P(k'|v, s, t, ; \hat{\mathbf{\Lambda}})} \tag{3.7}$$

混合比率は，道路セグメント s 及び時刻 t 固有のパラメタなので，該当部分のデータのみを用いて推定する．以上の方法により，全てのセグメントで用いる共通 GMM の平均，分散と各セグメントの道路状態を表す個別混合比率を得る．次に，このモデルを用いてデータの圧縮を行う．

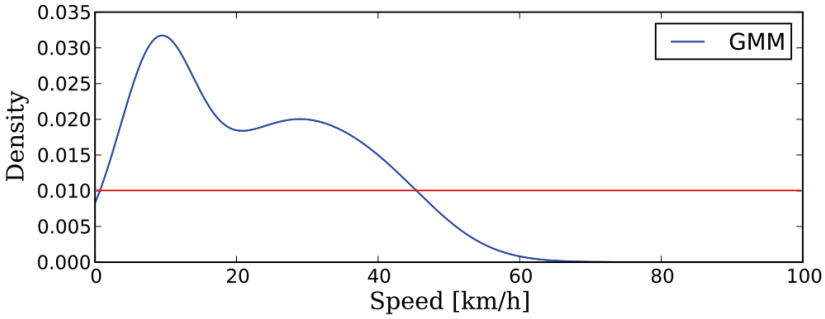

図 3.1: 圧縮 サンプル

3.4 データ圧縮

得られた個別 GMM を用いて，平常と思われるデータを破棄し，データの圧縮を行う．推定された個別 GMM は平常状態を表すモデルとなっているため，このモデルから外れている特徴をもつデータは異常と見なす．ある閾値 α を設定し，モデルの確率密度とこの閾値を比べ，大きければ平常データ，小さければ異常データとする．特徴量 v における，モデルの確率密度は以下の式で表される．

$$F(v) = \sum_{k=1}^{K} \theta_{stk} \mathcal{N}(v|\mu_k, \Sigma_k)$$

この $F(v)$ について，$F(v) \geq \alpha$ となる特徴量 v をもつデータを平常データ，$F(v) < \alpha$ となる特徴量 v を持つデータは異常データとする．平常データはモデルと α を用いて再現できるとし，異常データのみを蓄積することでデータの圧縮を行う．図.3.1 に圧縮の図を示す．青の曲線が，正規分布の足しあわせで表されるモデルであり，赤の直線が閾値である．ここでは $\alpha = 0.01$ としている．この図において閾値直線を上回っている 45km/h あたりまでのデータは平常データとみなして破棄し，下回っているデータを異常とみなし蓄積する．これにより，データの大部分を破棄し圧縮を行う．

第 4 章
実験データ

図 4.1: 神保町周辺地図

本研究で用いるプローブカーデータについて述べる．プローブカーデータとは，センサーが搭載された車から得られるデータであり，速度，位置情報，時刻など，搭載車の様々な状態を常時得ることができる．搭載車はこの数年で急速に増加しており，これからも増加することが予想される．

このプローブカーデータについて，東京都千代田区内，専大前交差点から神保町交差点を抜けて，駿河台下交差点にかけてのデータについて集計した．対象部分を図.4.1 示す．この区域のデータについて，時間を 1 時間ごとに区切り速度分布と平均速度を算出した．データは 2012 年 2 月 1 日から 2012 年 12 月 31 日までの全データを用いた．

表 4.1 に示すのが時間帯ごとの平均速度である．日中の速度は小さく，夜間は大きくなっている．これは夜間の方が交通量が減少し，道路が空くためと考えられる．また、速度分布には停車しているデータが大量に含まれており分布が偏るため，停車していると判断できる速度 0, 1 を除いたものも同時に示す．速度 0, 1 を除いた時の平均速度は，日中では倍以上速度が大きくなるが，夜間ではそこまで大きくはならない．このことから，夜間は日中に比べ，停車データが少ないことがわかる．0, 1 を除いた平均速度を図にしたものが Fig.4.2 である．日中は 26km/h 前後であるが、夜間では 32km/h 前後で推移しており，これは日中に靖国通りが混雑するためであると考えられる．図.4.3 から図.4.26 に速度 0, 1 を除いた時間帯毎の速度分布図を示す．夜間の分布は低速部分と高速部分の差が大きくなっているが，速度分布は日中と夜間に大きな差は見られない．以下では，このデータを用いて神保町付近の道路状態を推定し，データ圧縮を行う．

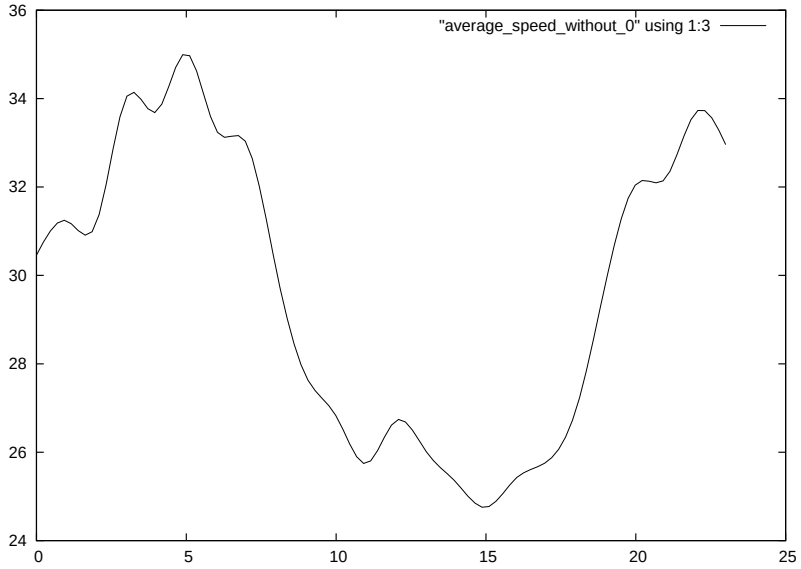

図 4.2: average speed without 0, 1 around jinbocho

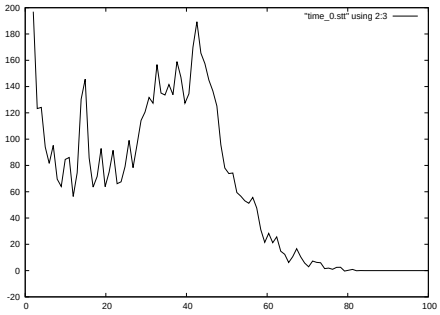

図 4.3: 平均速度 0:00-1:00

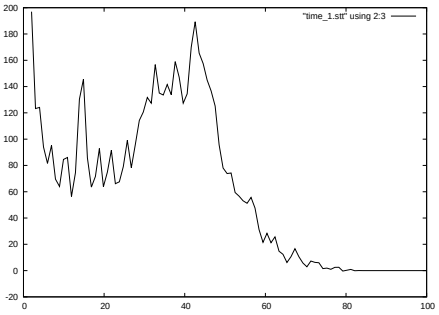

図 4.4: 平均速度 1:00-2:00

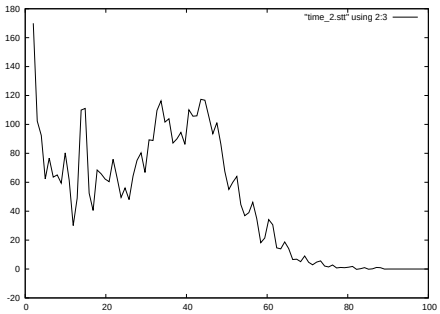

図 4.5: 平均速度 2:00-3:00

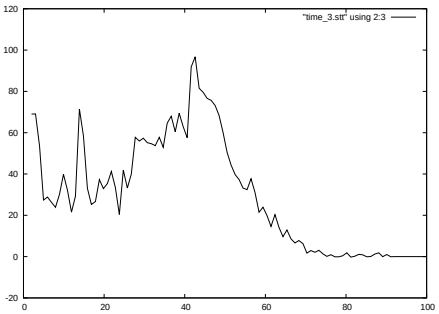

図 4.6: 平均速度 3:00-4:00

表 4.1: average speed around jinbocho

time	average speed	average speed (without 0, 1)
0	27.7	30.46
1	15.37	31.24
2	16.44	31.18
3	19.92	34.03
4	19.2	33.7
5	19.03	35.03
6	19.92	33.28
7	18.37	33.0
8	16.18	30.13
9	13.82	27.71
10	12.97	26.82
11	12.63	25.74
12	13.17	26.72
13	12.57	26.02
14	12.19	25.32
15	11.76	24.75
16	12.17	25.41
17	12.37	25.77
18	12.88	26.95
19	15.37	29.85
20	16.79	32.06
21	16.07	32.2
22	19.03	33.69
23	16.64	32.96

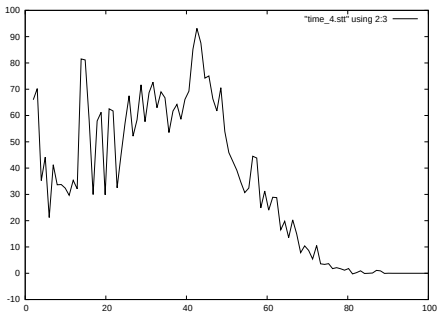

図 4.7: 平均速度 4:00-5:00

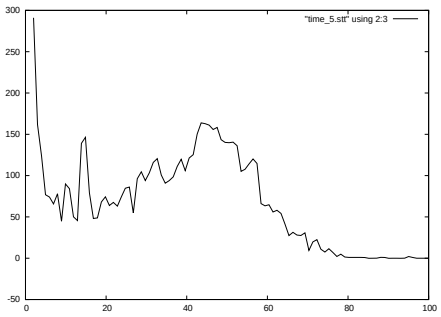

図 4.8: 平均速度 5:00-6:00

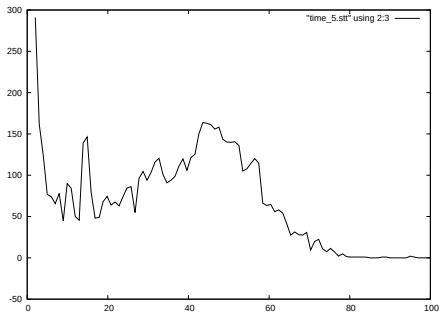

図 4.9: 平均速度 6:00-7:00

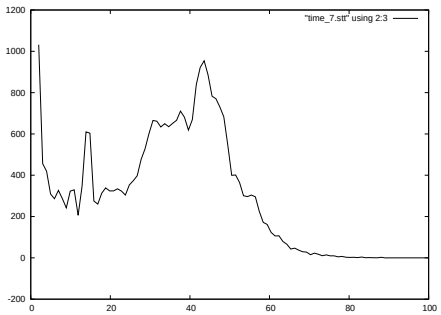

図 4.10: 平均速度 7:00-8:00

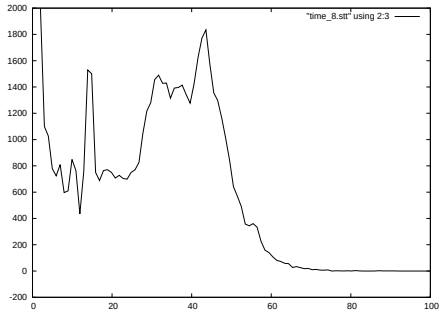

図 4.11: 平均速度 8:00-9:00

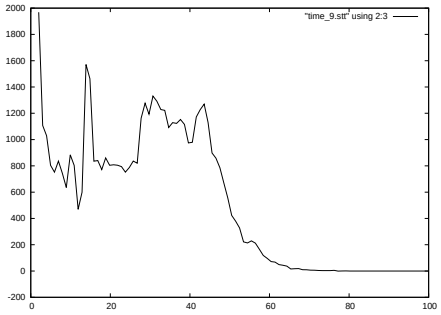

図 4.12: 平均速度 9:00-10:00

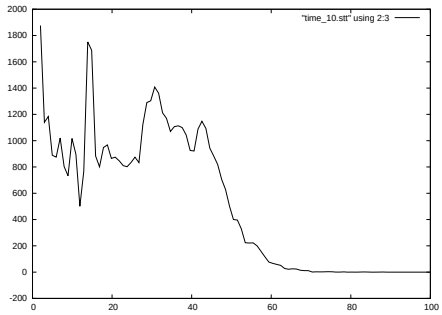

図 4.13: 平均速度 10:00-11:00

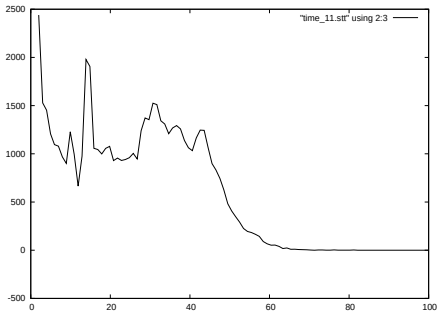

図 4.14: 平均速度 11:00-12:00

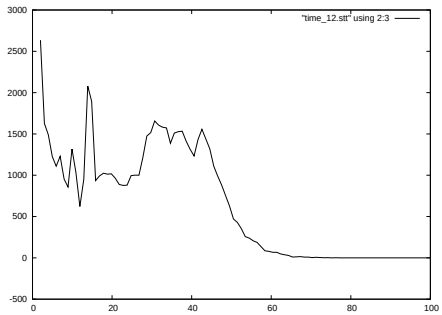

図 4.15: 平均速度 12:00-13:00

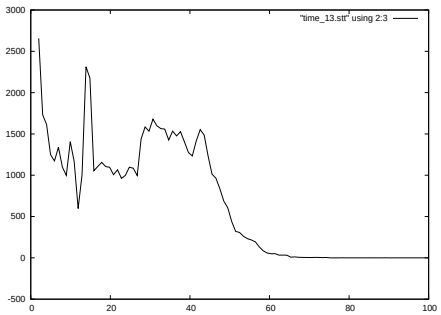

図 4.16: 平均速度 13:00-14:00

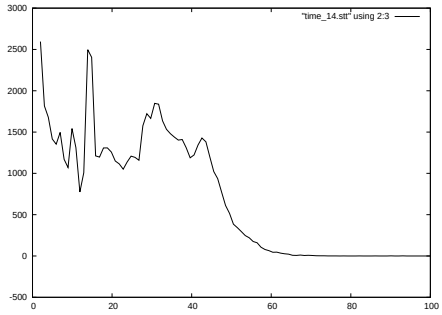

図 4.17: 平均速度 14:00-15:00

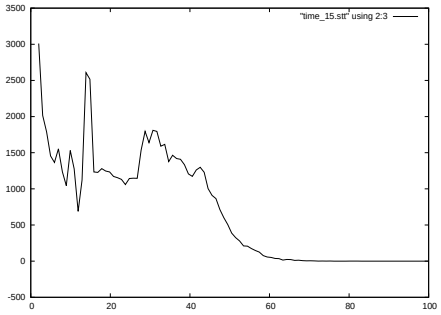

図 4.18: 平均速度 15:00-16:00

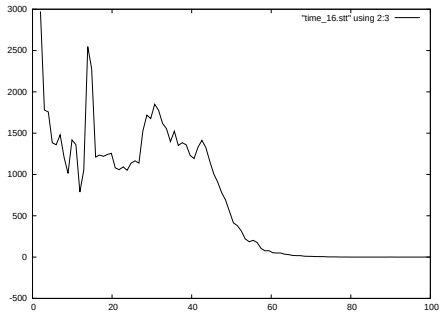

図 4.19: 平均速度 16:00-17:00

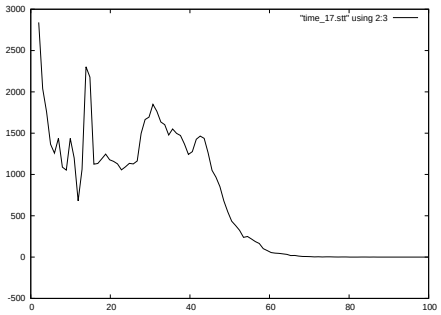

図 4.20: 平均速度 17:00-18:00

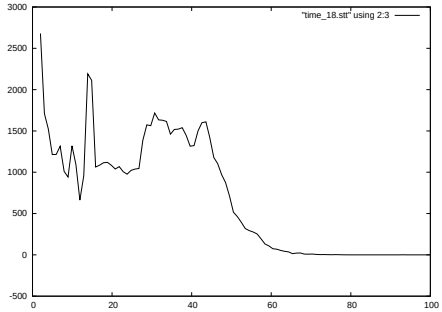

図 4.21: 平均速度 18:00-19:00

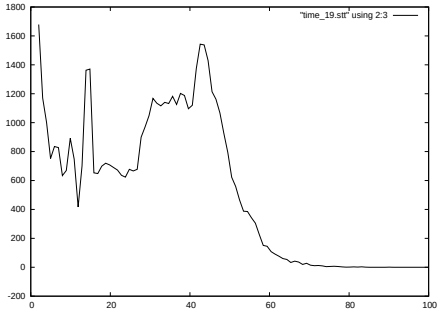

図 4.22: 平均速度 19:00-20:00

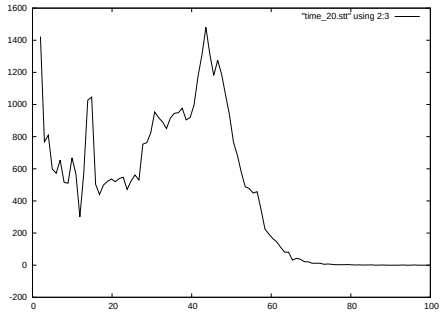

図 4.23: 平均速度 20:00-21:00

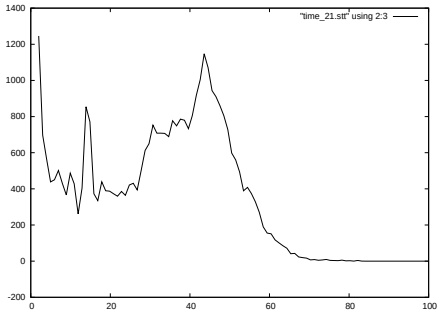

図 4.24: 平均速度 21:00-22:00

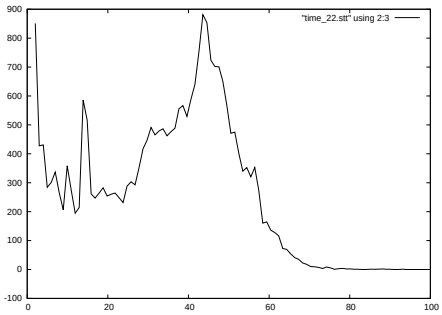

図 4.25: 平均速度 22:00-23:00

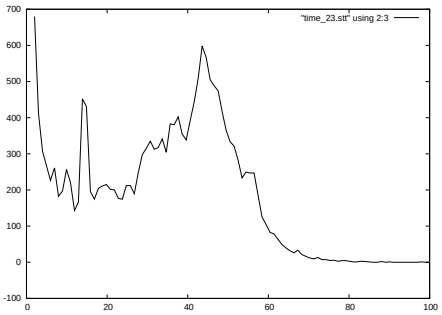

図 4.26: 平均速度 23:00-24:00

第 5 章
評価実験

図 5.1: 神保町セグメント

5.1 はじめに

アルゴリズムの有効性を示すため，実験を行った．実験は大きく分けて 3 つのステップからなる．

- モデル混合数の評価
- パラメタ推定と評価
- データ圧縮と評価

まずモデル混合数の評価では，正規混合分布の混合数を決定する．学習を行ったモデルとテストデータとの KL-Divergence を用いて，混合数ごとのモデルの評価を行い，適切な混合数を決定した．次のパラメタ推定と評価では，決定した混合数でのパラメタ推定を行い，テストデータと比較し適応しているかを確認する．最後に推定されたモデルを用いて，データ圧縮と評価を行う．データ圧縮では，基準となる閾値を $0.02, 0.01, 0.005$ と変化させた時の圧縮率と，再現率を評価した．評価では，閾値ごとにテストデータと再現データの図とともに，KL-Divergence も算出し評価した．

用いたデータは，2011 年 2 月 1 日から 2011 年 12 月 31 日までの，東京都神保町付近で得られたプローブカーデータである．このうち停車していると思われる，速度が 0 または 1 のデータを除いた約 90 万件を用いた．神保町付近を 50m ごとに区切り，全部で 15 のセグメントに分割し，それぞれのセグメントに対して 1 時間ごとに区切り 24 の時間帯に分けた．得られたデータのうち，速度のみを特徴量とし利用し 1 次元特徴量として実験を行った．90 万件のデータのうち，80 万件を学習データセット，10 万件をテストデータセットとし，モデルの混合数を $2, 4, 8, 16$ に設定してそれぞれ実験を行った．

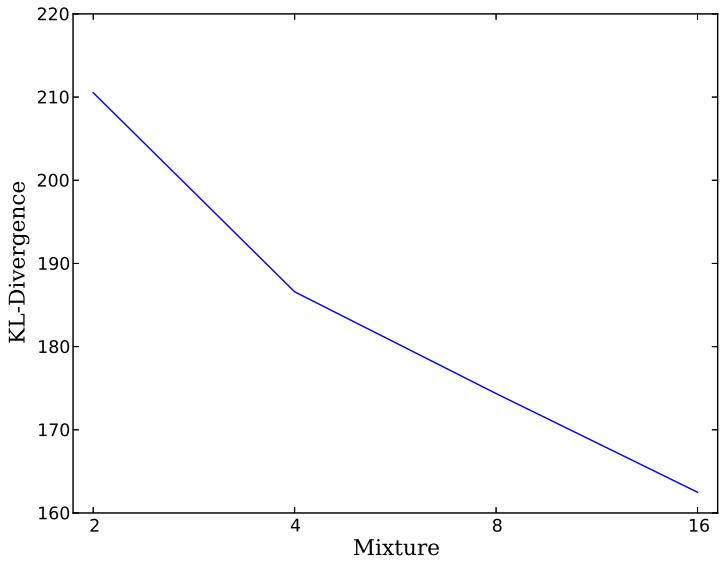

図 5.2: 混合数を変化させた時の KL-Divergence

5.2 混合数の評価

推定されたモデルと本来のデータとの差を計る指標として，KL-Divergence を用いた．分布 $Q(v)$ を GMM で推定を行った分布 $\sum_k \theta_{stk} N(v|\mu_k, \Sigma_k)$ とし，分布 $P(v)$ をテストデータセットの速度分布とする．ここで $Q(v)$ について，80 万件の学習データを用いて学習したモデル $\sum_{k \in K} \theta_{stk} \mathcal{N}(v|\mu_k, \Sigma_k)$ を用いた．各セグメントごとに KL-Divergence を算出し，全ての和を取った．

$$\sum_{s \in S} \sum_{t \in T} \sum_{v \in \boldsymbol{d}_{st}} P(v) \log \frac{P(v)}{\sum_{k \in K} \theta_{stk} \mathcal{N}(v|\mu_k, \Sigma_k)} \tag{5.1}$$

混合数を変化させた時の KL-Divergence の和を比較したものが図.5.2 である．混合数が大きくなる程 $P(v)$ と $Q(v)$ との距離は小さくなっており，実データに適応しているが，今回は混合数を 4 に設定して実験を行った．

5.3 EM アルゴリズムを用いたパラメタ推定

混合数 4 で EM アルゴリズムを用いてパラメタ推定を行った時の，尤度推移を図.5.3 に示す．図より，1 度目のパラメタ更新で大きく尤度が上がっていることが

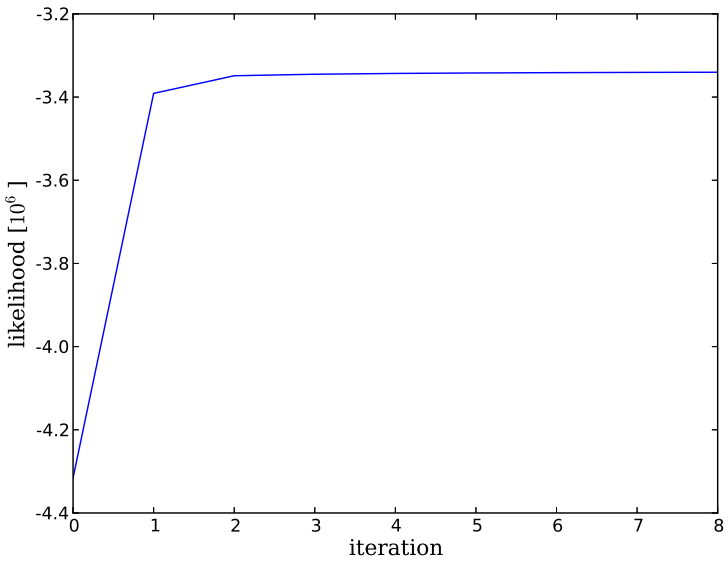

図 5.3: likelihood GMM 4 Mixture

読み取れる．混合数 4 の実験で推定された正規分布の平均と分散が図.5.4 である．セグメント，時間帯で共通の正規分布を用いることにより，それぞれのガウシアンが，平均が低い方から’混雑’，’低速’，’通常’，’高速’に対応すると見なすことができる．セグメントと時間帯ごとに分割されたデータを，これらの正規分布の重み付き和で表現した時の各混合比率が道路状況を表す．例えば，混雑，低速を表す正規分布の重みが大きければ道路が混雑しており，渋滞していると考えられるという具合である．

推定された共通 GMM を，各セグメントに適応する．データは時間帯毎に分かれているがこれは，同じセグメントであっても時間帯による道路状況の違いがあると考え，それを捉えるためである．図.5.5 は，神保町の第 10 セグメントにおける昼間の混雑時の速度分布である．このセグメントは，信号と距離があり，昼間でも比較的スムーズに車が流れるセグメントとなっている．これに対し，同じセグメントにおける非混雑時での速度分布が図.5.6 である．図より非混雑時は速度が大きくなっており，車の流れが速いことがわかる．昼間の混雑時のデータを正規混合分布で表現したものが図.5.7 である．昼間の混雑する時間帯では，30km あたりの正規分布の重みが大きくなっており，45km 辺りに平均をもつ正規分布の重みはほとんどない混合比率となっている．それに対し，非混雑時での速度分布における混合比率を図.5.8 に示す．混雑時のデータではほとんど重みのなかった速度

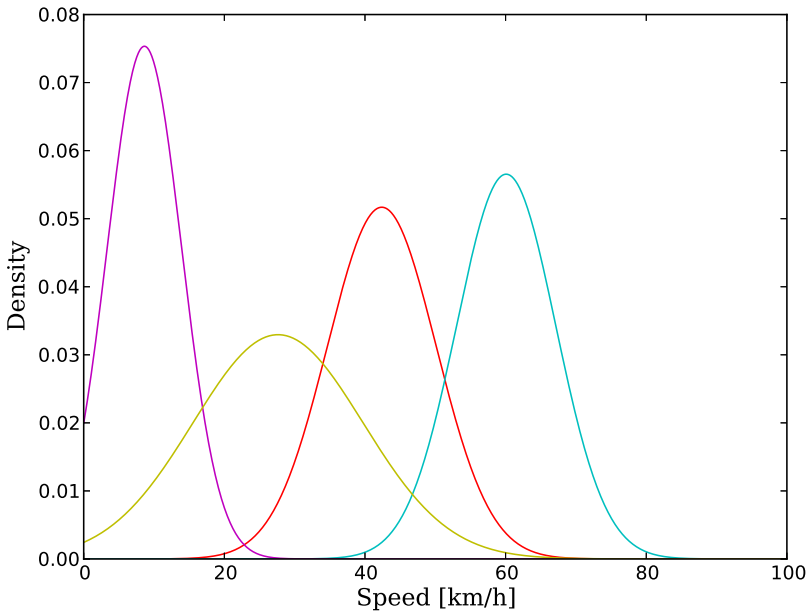

図 5.4: Gaussian 4 Mixture

45km あたりの正規分布の重みが大きくなっており，車の流れが速い非混雑時の道路状況をうまく表現している．このように共通 GMM でデータを表現することで，その混合比率が道路状況を表す特徴量となっている．

5.4 データの圧縮

次に，この推定された GMM を用いてデータの圧縮を行った．推定された個別 GMM は各セグメントにおける平常状態を表すモデルである．このモデルで表すことのできるデータは，蓄積する必要がないため破棄する．全テストデータに対し，閾値 α をそれぞれ 0.02, 0.01, 0.005 に設定した時の破棄されたデータ数と，その比率を示したのが表 5.1 の 2 行目である．閾値 0.02 で約半分，0.01 では 9 割を超えるデータの圧縮を実現した．表 5.1 の 3 行目に，図.5.7 のデータに対し，同様の閾値に設定した時の破棄されたデータ数と，その比率を示した．閾値 0.02 で約 4 割，0.01 では 9 割を超えるデータの圧縮を実現した．同様に図.5.8 のデータに対し，同様の閾値をに設定した時の破棄されたデータ数と，その比率を表 5.1 の 4 行目に示した．このデータに関しては閾値 0.02 で約 6 割，0.01 で約 8 割のデータ圧縮率となった．3 行目と比較すると，閾値 0.02 では 41% と 63%，0.01 では 96% と

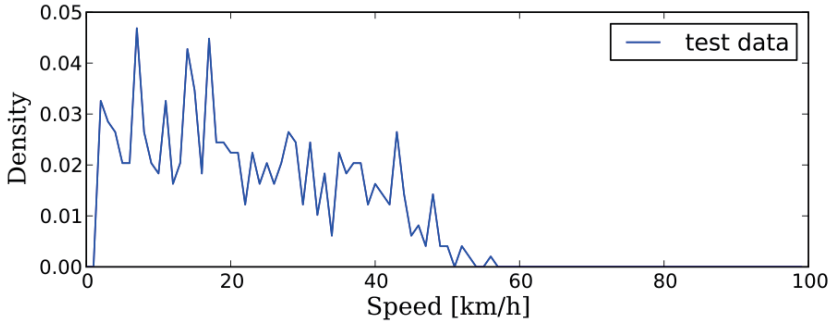

図 5.5: 神保町 混雑時

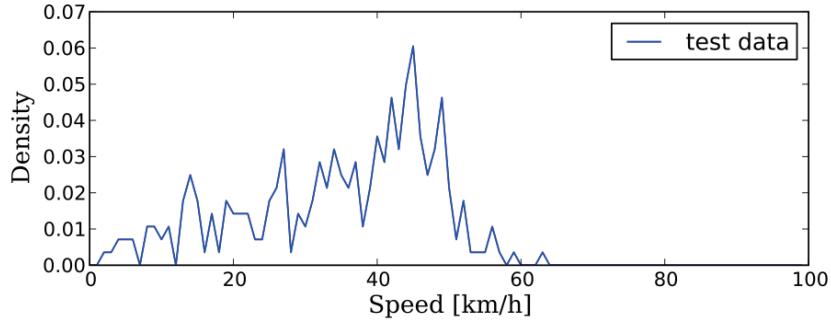

図 5.6: 神保町 非混雑時

83% など，分布によってばらつきがあることがわかる．図.5.7 のように広く分布している場合，閾値が一定以上小さければ，その部分が広く圧縮されるため効果が大きい．対して図.5.8 のように一部分に大きく分布している場合は，その部分のみの圧縮になるため前者に比べ効果が小さい．しかし，大部分のデータがそこに集中しているため十分な圧縮が行われている．

大部分を占める平常状態のデータを破棄し大幅な圧縮を行ったが，蓄積しているデータから元データを再現する必要があるため，再現性の実験を行った．蓄積している情報は，圧縮時の閾値 α，α を下回った異常データ，平常状態モデル，破棄データ数である．これらの情報から，再現データを生成する．平常状態モデルにおいて α を上回っている部分のみを取り出し，正規化した再現分布 $Q'(v)$ を作る．

$$Q'(v) = \begin{cases} 0 & (Q(v) < \alpha) \\ Q(v)/S_{st\alpha} & (Q(v) \geq \alpha) \end{cases} \tag{5.2}$$

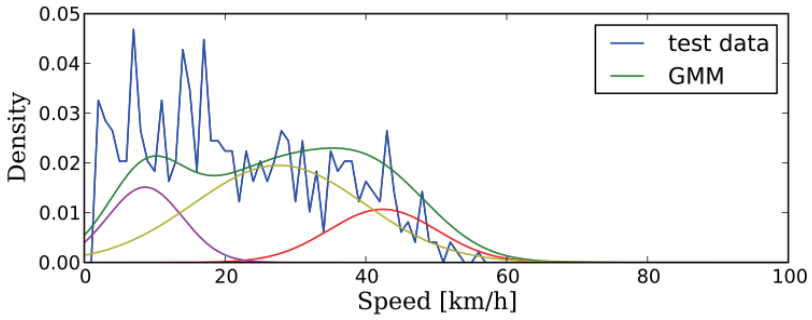

図 5.7: 神保町 混雑時 GMM

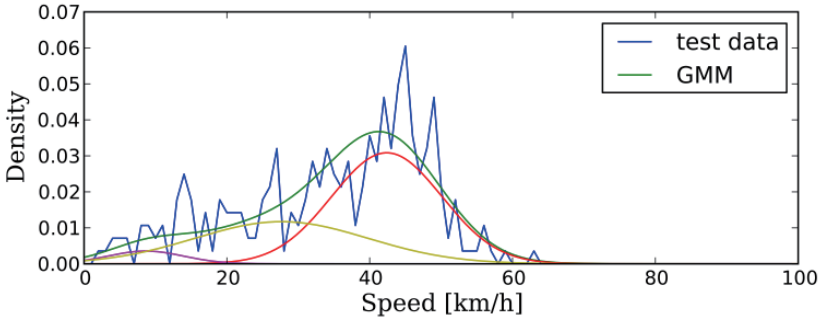

図 5.8: 神保町 非混雑時 GMM

ただし，

$$S_{sta\alpha} = \sum^{V} Q(v) \quad (Q(v) \geq \alpha)$$

とする．この $Q'(v)$ から，破棄データ数だけランダムサンプリングを行う．こうして得られたデータを平常データとし，蓄積している異常データと合わせて再現データとする．この方法により得られた図.5.7 に対する $\alpha = 0.005$ での再現データが図.5.9 である．$\alpha = 0.005$ では圧縮率が 99% を超えており，実データはほとんどないため，元データと比べると分布の違いが目立つところが散見される．特に，40km/h 辺りに確率密度が小さくなっているところがあり，再現が粗くなっていることがわかる．$\alpha = 0.01$ での再現データが図.5.11 である．こちらは圧縮率が約 96% であり，$\alpha = 0.005$ と比べ蓄積している実データが多く，元データを再現できている．細かい分布の再現はできていないが，$\alpha = 0.01$ でおよそ 96% の圧縮率があることを考慮すると，十分な再現ができているといえる．$\alpha = 0.02$ での再現データが図.5.13 である．こちらは圧縮率が約 41% であり，半分以上の実データ

表 5.1: データ圧縮の結果

	data	0.02	0.01	0.005
全データ	108371	52946	100296	106084
(圧縮率)		48.9%	92.5%	97.9%
神保町混雑時	491	203	470	487
(圧縮率)		41.3%	95.7%	99.2%
神保町非混雑時	281	177	233	277
(圧縮率)		63.0%	82.9%	98.5%

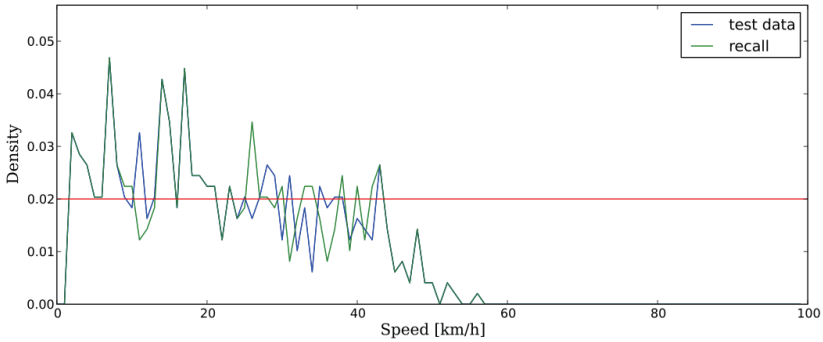

図 5.9: 神保町 混雑時の $\alpha = 0.02$ での再現データ

を蓄積しているため，元データに近い再現ができている．図.5.15 に，それぞれの α での再現データと元データとの KL-Divergence を算出した．α が大きくなるにつれ，再現性は高くなっていることがわかる．しかし，それに従って圧縮率は悪くなるため，扱うデータによって最適な α を設定する必要がある．

また図.5.8 に対する $\alpha = 0.005$ での再現データが図.5.10 である．$\alpha = 0.005$ では圧縮率は 98% であり，先ほどと同様に元データの分布との違いが目立つ部分が見られる．$\alpha = 0.01$ での再現データが図.5.12，$\alpha = 0.02$ での再現データが図.5.14 である．50km/h あたりのバラつきが再現できていないが，α が大きくなるほど再現率は高くなり，大まかな分布の再現はできているといえる．図.5.16 に，それぞれの α での再現データと元データとの KL-Divergence を算出した．図.5.15 と同様に，α が大きくなるにつれ，再現性は高くなっていることがわかる．しかし，それに従って圧縮率は悪くなるため，α を適切に設定する必要がある．また，α に関わらず，実データには分布のバラつきが見られるが，モデルではバラつきがなく滑らかになっているため，この部分の再現がうまくできていない．しかし，大まかな分布の再現はできているといえる．

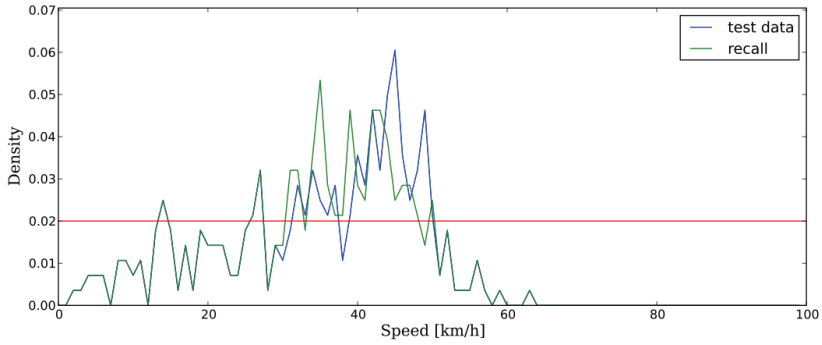

図 5.10: 神保町 非混雑時の $\alpha = 0.02$ での再現データ

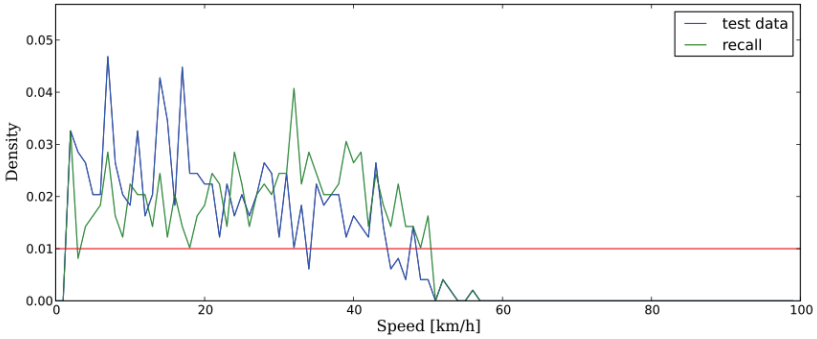

図 5.11: 神保町 混雑時の $\alpha = 0.01$ での再現データ

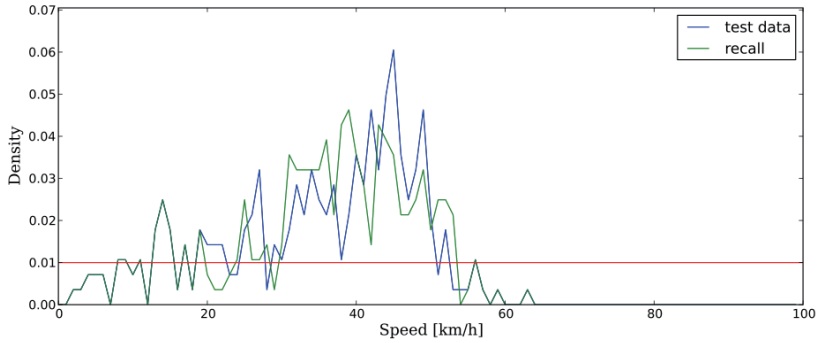

図 5.12: 神保町 非混雑時の $\alpha = 0.01$ での再現データ

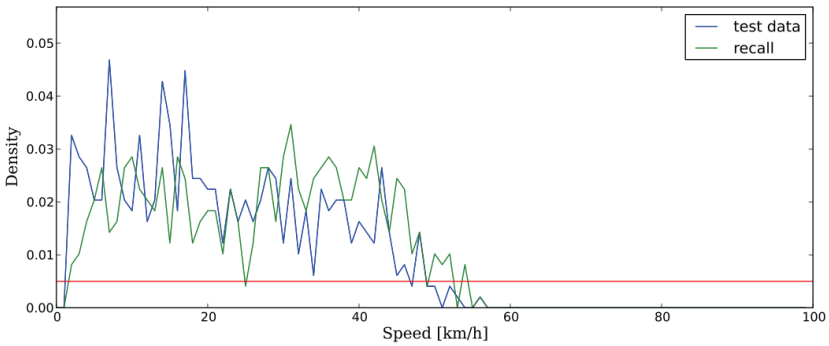

図 5.13: 神保町 混雑時の $\alpha = 0.005$ での再現データ

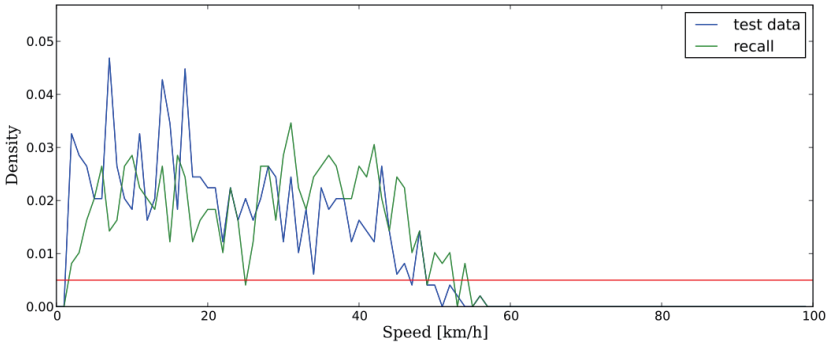

図 5.14: 神保町 非混雑時の $\alpha = 0.005$ での再現データ

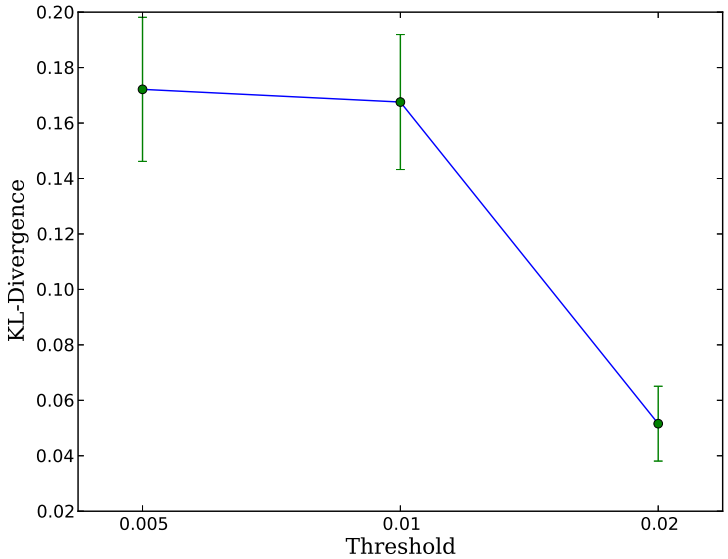

図 5.15: 神保町 混雑時の再現データの KL-Divergence

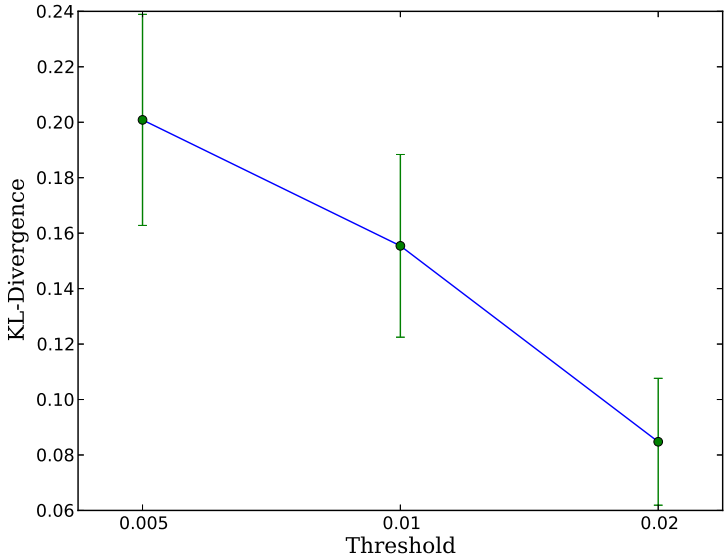

図 5.16: 神保町 非混雑時の再現データの KL-Divergence

第 6 章
結論

6.1 まとめ

本研究では，ビッグデータであるストリームデータについてデータ圧縮の研究を行った．ストリームデータの中でも大部分が平常状態データであり異常データが少ないものに注目し，これを蓄積するのに効果的な圧縮手法を提案した．

平常状態のデータを機械学習し，道路状態モデルで表すことで平常状態のデータを蓄積する必要をなくし，モデルから外れる異常データとモデルのみを蓄積した．これにより，蓄積データから元データへの再現性を確保しつつ，大幅な圧縮を達成した．モデルの学習は，EM アルゴリズムを用いて正規混合分布でモデル化した．その際データをセグメント，時間帯で分割することで，実際の道路状況に合わせた推定を行った．データの圧縮では閾値を定め，推定したモデルと照らし合わせて出現率の低いデータのみを蓄積し，それ以外を破棄することでデータを大きく圧縮した．出現率が 1% 以下の異常データだけを蓄積することで，90% を超える圧縮率を達成した．この圧縮の有効性を示すために再現実験を行い，比較的再現できていることを確認した．今回は 4 混合でのモデル化をし圧縮を行ったが，データのバラつきがモデルでは平滑化されており，十分に再現できていなかった．混合数を大きくすれば，データのバラつきを表現することが可能と考えられる．混合数を増やすことによるデメリットは学習に時間がかかることであり，これを考慮しつつ最適な混合数を検証していきたい．

6.2 今後の展望

また，今後の課題は大きく分けて以下の 3 つある．

- データの多次元化
- 効果的なデータ分割
- オンライン学習の導入

1 つ目はデータの多次元化である．今回の手法ではデータ特徴量を速度のみにしたが，多次元への拡張は容易なため，効果的な特徴量を考え多次元化を行いたい．今回のモデルでは，速度が異常でなければ急加減速や急旋回などの異常データは取得することができない．そのため，他に加速度や方向を特徴として考えており，加速度を追加することによって加減速の異常が，方向を追加することで旋回の異

常が取得できるものと見込んでいる．2つ目は効果的なデータ分割である．今回セグメント分割，時間分割に関しては手動で分割を行ったが，データに合わせて自動的に効果的な分割を行いたい．例えば，セグメントであれば，道路の幅が同じになるように区切ったり，交差点を全て含むように区切ることが望ましい．今回は1つの幹線道路を手動で区切ったため，理想的なセグメントの区切り方になっているが，同じような分割を自動で行える手法を考えたい．3つ目はオンライン学習の導入である．今回の手法では，想定外の出来事によりモデルと実際のデータ分布が異なった場合，対応することが難しい．実際のデータのピークが圧縮対象範囲に含まれていれば，分布が変わっているにも関わらず破棄されてしまうし，また対象外にピークがある場合は本来破棄されるデータが残ってしまうため，データ圧縮率が大きく低下する．ストリーム処理に適用することを考えると，オンライン学習によりモデルを更新していくことが効果的だと考えられるため，今後研究していきたい．

謝辞

本研究を行うにあたり，多くの方々にお世話になりました．

指導教員である安達淳教授には，最後まで絶え間ない教えを頂き，様々な知識に触れる機会や，恵まれた研究環境を与えてくださいました．深く感謝いたします．

高須淳宏教授には，修士研究について多くのご指導頂き，お陰様で修士研究をこのような形にすることができました．ありがとうございます．

相原健郎教授には，リサーチ・アシスタントとして面白う分野の研究に触れる機会を与えて下さいました．また，研究についても多くの助言を頂き，大変参考にさせていただきました．ありがとうございます．

安達研究室の皆様には公私ともに大変お世話になりました．秘書の久芳藍様には，NII での研究生活や，出張など様々面で支えとなってくださいました．OB の鈴木貴敦さんには，まだ至らない修士 1 年生の時に大変お世話になりました．楽しい研究生活を作って頂いた ChuYimin さん，木村光樹さん，赤塚裕人君，木下僚君，ありがとうございます．大変お世話になりました．

参考文献

[ATA10] Takahashi Akira, Atsuhiro Takasu, and Jun Adachi. Language model combination for community-based *q&a* retrieval. *International Conference on Tools with Artificial Intelligence*, pp. 241–248, 2010.

[BBA11] Chui Michael Bughin Jacques Dobbs Richard Roxburgh Charles Brown Brad, Manyika James and Hung Byers Angela. Big data:the next frontier for innovation,competition, and productivity. May 2011.

[Bis06] Chiristopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[BL09] David M. Blei and John D. Lafferty. Visualizing topics with multi-word expressions. 2009.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, 2003.

[BPPM93] Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, pp. 263–311, 1993.

[DCR+08] Fabritiis De, Corrado, Ragona, Roberto, Valenti, and Gaetano. Traffic estimation and prediction based on real time floating car data. *2008 11th International IEEE Conference on Intelligent Transportation Systems*, Vol. 24, No. 1, pp. 197–203, 2008.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, Vol. 39, No. 1, pp. 1–38, 1977.

[FDG08] Ragona Roberto Valenti Fabritiis De, Corrado and Gaetano. Traffic estimation and prediction based on real time floating car data. *11th International IEEE Conference on Intelligent Transportation Systems*, No. 1, pp. 197–203, 2008.

[Fuh92] Norbert Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, Vol. 35, pp. 243–255, 1992.

[JH07] P. Olsen J. Hershey. Approximating the kullback leibler divergence between gaussian mixture models. *Proc. ICASSP*, Vol. 4, pp. 317–320, 2007.

[JYL07] Brian Noble Jungkeun Yoon and Mingyan Liu. Surface street traffic estimation. *In Proceedings of the 5th international conference on Mobile systems, applica- tions and services*, pp. 220–232, 2007.

[KL51] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86, 1951.

[Kul68] Solomon Kullback. Information theory and statistics. *Dover PublicationsInc*, 1968.

[LM05] Y. Li and M. Mcdonald. Motorway incident detection using probe vehicles. *Proceedings of the ICE-Transport*, pp. 11–15, 2005.

[XCG10] Enxiang Liu Xueqing Cheng, Wenfang Lin and Dan Gu.

Highway traffic incident detection based on bpnn. *Procedia Engineering*, pp. 482–489, 2010.

[小西 04] 小西貞則, 北川源四郎. 情報量基準. 朝倉書店, 2004.

発表文献

[1] 山本敬介，高須淳宏，相原健郎，安達淳，”混合分布を用いた道路状態推定”，第 88 回人工知能基本問題研究会，2013.