# 数理情報工学特論第一 【機械学習とデータマイニング】 4章：教師なし学習①

かしま　ひさし
鹿島 久嗣
（数理 6 研）

kashima@mist.i.～

東京大学
THE UNIVERSITY OF TOKYO

# 教師なし学習の基本的な問題設定と、そのひとつである未観測値の推定を学びます

- 教師なし学習とは何か
- 多次元の確率分布
  - 多次元正規分布
  - 多項分布
- 教師なし学習の主要タスク
  - タスク1: 確率分布そのものを用いた分析
  - タスク2: データの確率評価
  - タスク3: 未観測値（欠損値）の推定
  - タスク4: 潜在変数の推定
- 多次元正規分布における未観測値の推定（タスク3）
  - 周辺分布
  - 条件付き分布

## 教師なし学習とは

# 教師なし学習は、データの従う確率分布の推定を目的とします

- 教師なし学習：入力$x$の従う（多次元）確率分布P($x$)の推定

  ⇔教師つき学習：入力$x$に対する出力$y$の条件付き確率P($y|x$)を推定

- 教師なし学習においては、与えられた（訓練）データ集合において出力に関する情報（教師信号）が与えられることはない
  - （訓練）データ集合としては$\{ x^{(i)}\}_{i=1}^{N}$ のように、$N$個の（教師つき学習でいうところの）入力のみが与えられる
  - 入力のみからモデルを推定するため、教師なし学習と呼ばれる
- ただし、教師なし学習でも、**潜在変数**と呼ばれる、出力 $y$ に対応する変数が用いられることもある
  - これらはデータ集合において直接的に観測されることはなく、あくまで仮想的な変数として用いられる

# 教師なし学習には、高次元空間上での確率分布を推定するという難しさがあります

- 教師つき学習では：
  - 入力の特徴ベクトル$\phi(x)$は多次元($D$次元)
  - 出力は主に1次元（もしくは比較的小さな次元）
- つまり、教師つき学習での（条件付き）確率分布は1次元空間上での確率分布を考えていたのであった
- 一方で、教師なし学習では$\phi(x)$が分布する多次元空間（$D$次元）における確率分布を扱うことになる
- 高次元の確率分布を扱う必要があるという点で、教師つき学習とは別の難しさがある

# 多次元の確率分布

# 代表的な分布2つ：多次元正規分布と多項分布

- まずは$P(x)$のモデルとして、どのような確率分布がよく用いられるかを見ることにする
- 教師なし学習におけるモデルは「データ（特徴ベクトル）を生成するモデル」である
- 扱う特徴ベクトルの種類や性質に応じて、適切なモデルを選ぶ必要がある
  - 実数値特徴ベクトルを扱う代表的モデル
    - 多次元正規分布
  - 離散値特徴ベクトルを扱う代表的モデル
    - 多項分布

# 正規分布は実数値を取る変数の最も代表的な確率モデルです

- データの特徴ベクトルの次元が1次元の場合（$\boldsymbol{\phi}(x) = \phi(x) \in \mathcal{R}$）の正規分布の確率密度関数：

$$P(\phi(x); \mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\phi(x)-\mu)^2}{2\sigma^2}\right)$$

  - $\mu$ ：平均パラメータ
  - $\sigma^2$ ：分散（$\sigma$は標準偏差）パラメータ
- 確率 $P(\phi(x); \mu, \sigma^2)$ は $\phi(x) = \mu$ のときに最大値をとり、一方 $(\phi(x) - \mu)^2$ が大きくなるに従い小さくなっていくような単峰性の分布である

# （一次元）正規分布を多次元に拡張したものが多次元正規分布です

- 多次元正規分布の確率密度関数：

$$P(\boldsymbol{\phi}(x);\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})\right)$$

  - $\boldsymbol{\mu}$：$D$次元の平均ベクトル（各要素が特徴ベクトルの各次元の平均）
  - $\boldsymbol{\Sigma}$：分散共分散行列（$D \times D$ の対称な正定行列）
  - $|\boldsymbol{\Sigma}|$：行列$\boldsymbol{\Sigma}$の行列式
- $\boldsymbol{\Sigma}$の非対角成分は特徴ベクトルの要素間の相関関係を表す
- $\boldsymbol{\Sigma} = \sigma\mathbf{I}$ と置くと、これは要素間の相関がまったくない状態を表し、密度関数は独立な1次元の正規分布の同時確率（確率の積）：

$$P(\boldsymbol{\phi}(x);\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{d=1}^{D} P(\phi_d(x);\mu_d,\sigma^2)$$

# 多次元正規分布の分散共分散行列を精度行列で置き換えた表現もあります

- 密度関数の分散共分散行列の逆行列を $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ と置き換えてみると多次元正規分布の別表記が得られます

$$P(\boldsymbol{\phi}(x); \boldsymbol{\mu}, \boldsymbol{\Lambda}) \equiv \frac{1}{(2\pi)^{D/2}} |\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\left( -\frac{1}{2} (\boldsymbol{\phi}(x) - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\boldsymbol{\phi}(x) - \boldsymbol{\mu}) \right)$$

  - $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ ： 精度行列（Σと同じく対称な正定値行列）

- 精度行列は、特徴ベクトルの要素間の直接的な依存関係を表す

# 精度行列は、特徴ベクトルの要素間の直接的な依存関係を表しています

- $\bar{\phi}(x) \equiv \phi(x) - \boldsymbol{\mu}$ と置くと、密度関数の指数部分は：

$$-\frac{1}{2}\bar{\phi}(x)^\top \boldsymbol{\Lambda} \bar{\phi}(x) = -\frac{1}{2}\sum_{k,\ell=1}^{D} [\boldsymbol{\Lambda}]_{k,\ell}\bar{\phi}_k(x)\bar{\phi}_\ell(x)$$

- 精度行列 $\boldsymbol{\Lambda}$ の $(k,l)$-要素 $[\boldsymbol{\Lambda}]_{k,l}$ が $k$ 番目の要素 $\bar{\phi}_k(x)$ と $l$ 番目の要素 $\bar{\phi}_l(x)$との「絡み」の程度を表していることがわかる。
- $[\boldsymbol{\Lambda}]_{k,l} = 0$ のときには$\bar{\phi}_k(x)\bar{\phi}_\ell(x)$の項が存在しないため、$k$ 番目の要素と $l$ 番目の要素の間には直接的な関係がないということになる
- $\boldsymbol{\Lambda}$ は単に密度関数を決定するパラメータであるだけでなく、それ自体が、変数間の関係の構造という重要な定性的情報をもつ
  - 後に、精度行列に対する$L_1$正則化を用いて多次元正規分布を推定（グラフィカルラッソ）する際に改めて触れる

# 正規分布は0以上の整数値を取る変数の最も代表的な確率モデルです

- 多項分布：離散値、とくに0以上の整数値を取るような特徴ベクトルを扱うための確率分布
  - 例えば、文書のbag-of-words表現など、各特徴が何らかの「個数」を表現する場合が多くの場面で現れる
- 多項分布の確率密度関数：

$T$ 個の対象を $\phi_1(x), \ldots, \phi_D(x)$ 個ずつに分割する場合の数

$$P(\boldsymbol{\phi}(x); \boldsymbol{\mu}, T) \equiv \binom{T}{\phi_1(x), \phi_2(x), \ldots, \phi_D(x)} \prod_{d=1}^{D} \mu_d^{\phi_d(x)}$$

  - $\boldsymbol{\mu}$：確率の制約をみたす$D$次元ベクトル

$$\sum_{d=1}^{D} \mu_d = 1, \qquad \mu_d \geq 0$$

  - $T$：自然数

$$\sum_{d=1}^{D} \phi_d(x) = T$$

# 多項分布は、不均一なD面体さいころによる各面の出現回数の分布として理解できます

- 多項分布は、不均一な$D$面体さいころのイメージで理解できる
  - $\mu$の各次元は、このさいころを1回振った時
  - $T$ はこのさいころを何回振るかという回数
  - $P(\boldsymbol{\phi}(x)\,;\,\boldsymbol{\mu}, T)$ は、このさいころを $T$ 回振り終わった時に、各面が出た回数が $\boldsymbol{\phi}(x) = (\ \phi_1(x), \ldots, \phi_D(x)\ )$ である確率を表す
- 例として、先ほどのbag-of-words表現を考えると、各単語の出現回数が $\boldsymbol{\phi}(x) = (\ \phi_1(x), \ldots, \phi_D(x)\ )$となる確率が $P(\boldsymbol{\phi}(x)\,;\,\boldsymbol{\mu}, T)$ である
  - $D$ 個の単語を含む単語リストがあり、その$d$番目の単語が出現する確率を$\mu_d$とする
  - 観測された文書の長さ（文書の含む単語数）が $T$ であるとする
  - 文書に含まれる各単語はお互いに独立に確率分布 $\boldsymbol{\mu}$ に従って出現すると仮定する（本当は必ずしも独立ではないはずだが）

## 多項分布では、要素の出現順は考慮していないことに注意する必要があります

- なお、多項分布の確率は文書中の単語の出現位置は考慮せず、単に「各単語の出現頻度」のみを扱っていることに注意する必要がある
- 文書を単に「単語を袋に詰めたもの」ではなく「単語の列」であると考えた場合、特徴ベクトルは

$$\boldsymbol{\phi}(x) \equiv (\phi_{1,1}(x), \ldots, \phi_{D,T}(x))$$

- のように、単語リスト$d$番目の単語が、文書中の位置$t$に出現した回数（この場合0か1になる）$\phi_{d,t}(x)$を用いて定義される
- そして、ある特定の単語列であるところの文書が出現する確率は：

$$P(\boldsymbol{\phi}(x); \boldsymbol{\mu}, T) \equiv \prod_{t=1}^{T} \prod_{d=1}^{D} \mu_d^{\phi_{d,t}(x)}$$

# 教師なし学習の主要タスク

# 教師なし学習の主要タスクは4つあります

- 教師なし学習においては通常、データ上の確率分布$P(\phi(x))$を何らかの形で推定することが行われるが、その使い道としては主に以下の4つが挙げられる
  - タスク1: 確率分布そのものを用いた分析
  - タスク2: データの確率評価
  - タスク3: 未観測値（欠損値）の推定
  - タスク4: 潜在変数の推定

# タスク1: 確率分布そのものを用いた分析<br>パラメータを見ることでデータの振る舞いを知ることができます

- パラメータ（によって決まる確率分布$P(\phi(x))$の形）を「鑑賞」することによって、データ全体がどのような形で分布しているかを知ることは、データについての知見を得るのに有効である
- 多次元正規分布の場合：
  - 分布の平均ベクトル$\boldsymbol{\mu}$と共分散行列$\boldsymbol{\Sigma}$を見ることで、データの全体的な振る舞いが見える（どのあたりの$\phi(x)$が最も確率が高く、また、そのまわりに$\phi(x)$がどのように散らばっているか）
  - 精度行列$\boldsymbol{\Lambda}$を見ることで、要素間の直接的な依存関係がわかる
    - $\phi(x)$ の各要素が遺伝子発現量（各遺伝子がどの程度働いているか）を表すとすると、精度行列の各要素は遺伝子間にどのような関係があるかを伺い知ることができる
    - 精度行列の要素が非零である場合、2つの遺伝子間には何からの密接な関係（協調関係や制御関係）があるものと期待できる。

# タスク1: 確率分布そのものを用いた分析 2つの分布を比較することでその差を検出できます

- 2つのデータ集合 $\mathcal{D} \equiv \{x^{(i)}\}_{i=1}^{N}$ および $\mathcal{D}' \equiv \{x'^{(i)}\}_{i=1}^{N'}$ （$x^{(i)}, x'^{(i)} \in \mathcal{X}$ ）から推定された2つの確率分布を $P(\phi(x))$ および $P'(\phi(x'))$ とする
- $P(\phi(x))$ と $P'(\phi(x'))$ が異なるか、異なるとしたらどのように異なるのかを調べることも実用上非常に重要
  - $\mathcal{D}$がある製造システムの正常時の挙動、$\mathcal{D}'$ は正常かどうかが疑わしいときの挙動とすると、これらの間に差があるということは、後者ではシステムに何らかの異常が起こっていると疑うことができる。
    - さらに、その差異を調べることで、どのような挙動の違いがあるのかを調べ、異常の原因を特定するのに役立てることができる
  - $\mathcal{D}$ を（bag-of-words表現などによって表現された）Webページなどの集合、$\mathcal{D}'$を最近作られたWebページの集合とし、その分布の差異を調べることで、最近現れた新規トピックの検出などを行うこともできる

# タスク2: データの確率評価<br>確率の小さいデータを異常データとして検出します

- 一旦 $P(\phi(x))$ が与えられれば、訓練データ集合中の任意のデータ、もしくは新たに入手したデータ$x$に対し、その確率値を評価できる
- つまり、そのデータ $x$ が確率分布 $P(\phi(x))$ から出てきたものであるとしたときに、$x$がどの程度出てきやすいものか、もしくは、どの程度出てきにくいものかを知ることができる
- データの確率評価が特に重要な役割をもつのが、外れ値検出
  - 外れ値検出とは、データ集合中で、他のデータとは大きく異なる「異常な」データを発見するタスク
  - 詐欺検知：大勢のユーザーのカード利用履歴中から不審な（通常行わない＝確率がきわめて低い）利用行動を取るユーザーを発見する
  - 雑音除去：計測機器により取得されたデータ集合から、突発的な雑音に乱された（通常観測されない＝確率がきわめて低い）値を見つけ取り除く

# タスク3: 未観測値（欠損値）の推定
# 一部の変数についての確率分布（周辺分布）を求めます

- 特徴ベクトル $\phi(x)$ =($\phi_1(x)$, …, $\phi_M(x)$, $\phi_{M+1}(x)$, …, $\phi_D(x)$) のうち（簡単のため最初の）$M$ 個の要素を纏めて $\phi_A(x) \equiv (\phi_1(x), \ldots, \phi_M(x))$ と書く
- また、残りの$D$-$M$個の要素を$\phi_B(x) \equiv (\phi_{M+1}(x), \ldots, \phi_D(x))$と書く
- このとき、特徴ベクトルの全要素についての同時分布 $P(\phi(x))$がわかると、これから、$\phi_A(x)$ についてのみの同時分布がわかる：

$$P(\phi_\mathsf{A}(x)) = \int P(\phi(x))\ \mathsf{d}\phi_\mathsf{B}(x) \qquad \text{（周辺化）}$$

- $\phi(x)$ の全要素ではなく、その一部 $\phi_A(x)$ にのみ興味がある場合に有効
  - 後で紹介する潜在変数モデル $P(x,y)$ は、実際には観測されない変数 $y \in \mathcal{Y}$ を含むモデルである
  - $\mathcal{Y} \equiv \{1,2,3\}$が3つの潜在的なカテゴリーを表すとすると、 $P(y)$ を知ることで、この潜在的なカテゴリーがそれぞれどの程度の割合で存在するのかを知ることができる

# タスク3: 未観測値（欠損値）の推定
# 一部の変数が分かった時の残りの変数を予測します

- 類似のケースとして、$\phi_{\mathrm{A}}(x) \equiv (\phi_1(x), \ldots, \phi_M(x))$ が観測されたとしたときの$\phi_{\mathrm{B}}(x) \equiv ( \phi_{M+1}(x), \ldots, \phi_D(x) )$ の条件付き確率分布

$$P(\phi_{\mathsf{B}}(x)|\phi_{\mathsf{A}}(x)) = \frac{P(\phi_{\mathsf{A}}(x), \phi_{\mathsf{B}}(x))}{\phi_{\mathsf{A}}(x)}$$

  を用いることで、$\phi(x) = (\phi_1(x), \ldots, \phi_M(x), \phi_{M+1}(x), \ldots, \phi_D(x))$ の要素のうちいくつかが観測された場合にその情報を利用して、残りの要素を予測することができる

- 応用例としては：
  - 住所や年収などといった顧客情報のうち一部が何らかの理由により欠損している場合に、残りの情報からこれを推測する

# タスク4: 潜在変数の推定
# 「隠れた情報」を表す仮想的な変数を考え、これを推定します

- 未観測値の推定と同じく、データ中に観測されない情報の推定であるが、まったく異なる状況として、潜在変数の推定がある
- このタスクは、興味のある変数が（訓練）データ集合においても観測されない潜在的な仮想変数であるという点で、未観測値の推定とは異なる
- つまり、$x$とは別にまったく観測されない変数$y$（潜在変数）と呼ばれる変数を仮定し、その条件付き分布 $P(y|x)$ を推定することが目的となる
  - $y$ は $x$ のもつ「隠れた情報」を表す
- たとえば、$y$を各データの属するグループであるとすると、P($y|x$) を用いてデータ集合を自動的にグループ分け（**クラスタリング**）することができる

多次元正規分布における未観測値の推定（タスク3）

# タスク3の未観測値（欠損値）の推定を、多次元正規分布を使って具体的にやってみます

- 特徴ベクトル $\phi(x)$ =($\phi_1(x)$, …, $\phi_M(x)$, $\phi_{M+1}(x)$, …, $\phi_D(x)$) のうち
  - 最初の$M$ 個の要素を $\phi_{\mathrm{A}}(x) \equiv (\phi_1(x), \ldots, \phi_M(x))$ と書く
  - 残りの$D$-$M$個の要素を $\phi_{\mathrm{B}}(x) \equiv ( \phi_{M+1}(x), \ldots, \phi_D(x) )$ と書く
- 特徴ベクトルの全要素についての同時分布 $P(\phi(x))$がわかったとき：
  - $\phi_{\mathrm{A}}(x)$ についてのみの同時分布（周辺分布）を求めたい
  - $\phi_{\mathrm{A}}(x)$ が観測されたとしたときの $\phi_{\mathrm{B}}(x)$ の条件付き確率分布をも求めたい

# 多次元正規分布に対して、一部の変数に対する（周辺）分布を、積分（周辺化）によって求めてみます

- $P(\phi(x))$ は（精度行列を用いて表現された）多次元正規分布：

$$P(\boldsymbol{\phi}(x);\boldsymbol{\mu},\boldsymbol{\Lambda}) \equiv \frac{1}{(2\pi)^{D/2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})\right)$$

- $\phi(x)$のうち最初の$M$個の要素 $\phi_A(x) \equiv (\phi_1(x), \ldots, \phi_M(x))$ についての同時分布は、残りの$D$-$M$ 個の要素 $\phi_B(x) \equiv (\phi_{M+1}(x), \ldots, \phi_D(x))$ による $P(\phi(x);\mu,\Lambda)$ の積分を求めることで得られる：

$$P(\boldsymbol{\phi}_{\mathsf{A}}(x)) = \int P(\boldsymbol{\phi}(x);\boldsymbol{\mu},\boldsymbol{\Lambda})\ \mathrm{d}\boldsymbol{\phi}_{\mathsf{B}}(x)$$

- つまり：

$$P(\boldsymbol{\phi}_{\mathsf{A}}(x)) = \frac{1}{(2\pi)^{D/2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \int \exp\left(-\frac{1}{2}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})\right)\ \mathrm{d}\boldsymbol{\phi}_{\mathsf{B}}(x)$$

# 精度行列を分割して、正規分布の指数部分を書き下します

- 平均$\boldsymbol{\mu}$の、最初の$M$要素の部分と残りの$D$-$M$要素の部分をそれぞれ$\boldsymbol{\mu}_{\mathrm{A}}$および$\boldsymbol{\mu}_{\mathrm{B}}$のように書き、$\bar{\phi}_{\mathsf{A}} \equiv \phi_{\mathsf{A}}(x) - \boldsymbol{\mu}_{\mathsf{A}}, \quad \bar{\phi}_{\mathsf{B}} \equiv \phi_{\mathsf{B}}(x) - \boldsymbol{\mu}_{\mathsf{B}}$とおくと：

$$\phi(x) - \boldsymbol{\mu} = \begin{bmatrix} \bar{\phi}_{\mathsf{A}} \\ \bar{\phi}_{\mathsf{B}} \end{bmatrix}$$

- $\boldsymbol{\Lambda}$の分割を以下のように置く：

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{AA} & \boldsymbol{\Lambda}_{AB} \\ \boldsymbol{\Lambda}_{BA} & \boldsymbol{\Lambda}_{BB} \end{bmatrix}$$

  - $\boldsymbol{\Lambda}_{\mathrm{AA}}$ ： $M \times M$行列、$\boldsymbol{\Lambda}_{\mathrm{BB}}$： $(D$-$M) \times (D$-$M)$行列
  - $\boldsymbol{\Lambda}$の対称性より、$\boldsymbol{\Lambda}_{\mathrm{AB}}^{\top} = \boldsymbol{\Lambda}_{\mathrm{BA}}$

- 多次元正規分布の指数部分は以下のように書き換えることができる：

$$-\frac{1}{2}(\phi(x)-\boldsymbol{\mu})^{\top}\boldsymbol{\Lambda}(\phi(x)-\boldsymbol{\mu}) = -\frac{1}{2}\left(\bar{\phi}_{\mathsf{A}}^{\top}\boldsymbol{\Lambda}_{AA}\bar{\phi}_{\mathsf{A}} + \bar{\phi}_{\mathsf{B}}^{\top}\boldsymbol{\Lambda}_{BB}\bar{\phi}_{\mathsf{B}} + 2\bar{\phi}_{\mathsf{B}}^{\top}\boldsymbol{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)$$

# 平方完成を用いて、積分に関わる部分と関わらない部分を分割します

- 以下の展開が成立する：

$$\left(\bar{\phi}_{\mathsf{B}} + \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)^\top \mathbf{\Lambda}_{BB}\left(\bar{\phi}_{\mathsf{B}} + \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right) = \bar{\phi}_{\mathsf{B}}^\top \mathbf{\Lambda}_{BB}\bar{\phi}_{\mathsf{B}} + 2\bar{\phi}_{\mathsf{B}}^\top \mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}} + \bar{\phi}_{\mathsf{A}}^\top \mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}$$

  – この式の右辺のはじめの2項が指数部分に出現する

- 指数部分をさらに以下のように書き換えることができる。

$$\begin{aligned}
&-\frac{1}{2}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})^\top \mathbf{\Lambda}(\boldsymbol{\phi}(x)-\boldsymbol{\mu}) \\
&= -\frac{1}{2}\Big(\bar{\phi}_{\mathsf{A}}^\top \mathbf{\Lambda}_{AA}\bar{\phi}_{\mathsf{A}} + \left(\bar{\phi}_{\mathsf{B}} + \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)^\top \mathbf{\Lambda}_{BB}\left(\bar{\phi}_{\mathsf{B}} + \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right) \\
&\qquad - \bar{\phi}_{\mathsf{A}}^\top \mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\Big) \\
&= -\frac{1}{2}\Big(\bar{\phi}_{\mathsf{A}}^\top(\mathbf{\Lambda}_{AA} - \mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA})\bar{\phi}_{\mathsf{A}} + \left(\bar{\phi}_{\mathsf{B}} + \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)^\top \mathbf{\Lambda}_{BB}\left(\bar{\phi}_{\mathsf{B}} + \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)\Big)
\end{aligned}$$

# 積分してみると、周辺分布が求まります

- 元々の式に戻すと以下のようになる：

$$P(\phi_\mathsf{A}(x)) = \frac{1}{(2\pi)^{D/2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}\int \exp\left(-\frac{1}{2}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{\phi}(x)-\boldsymbol{\mu})\right)\ \mathrm{d}\phi_\mathsf{B}(x)$$

$$= \frac{1}{(2\pi)^{D/2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\bar{\phi}_\mathsf{A}^\top(\boldsymbol{\Lambda}_{AA}-\boldsymbol{\Lambda}_{AB}\boldsymbol{\Lambda}_{BB}^{-1}\boldsymbol{\Lambda}_{BA})\bar{\phi}_\mathsf{A}\right)$$

$$\int \exp\left(-\frac{1}{2}\left(\bar{\phi}_\mathsf{B}+\boldsymbol{\Lambda}_{BB}^{-1}\boldsymbol{\Lambda}_{BA}\bar{\phi}_\mathsf{A}\right)^\top \boldsymbol{\Lambda}_{BB}\left(\bar{\phi}_\mathsf{B}+\boldsymbol{\Lambda}_{BB}^{-1}\boldsymbol{\Lambda}_{BA}\bar{\phi}_\mathsf{A}\right)\right)\ \mathrm{d}\phi_\mathsf{B}(x)$$

- 積分部分は、確率分布として（和が1になるように）正規化されていない正規分布の形をしているので、この部分が精度行列$\Lambda_{BB}$をもつ（正規化されていない）$D$-$M$ 次元の正規分布に対応していることに注意すると、この積分は $(2\pi)^{(D-M)/2}|\Lambda_{BB}|^{-1/2}$となるはずなので：

$$P(\phi_\mathsf{A}(x)) = \frac{1}{(2\pi)^{M/2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}|\boldsymbol{\Lambda}_{BB}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\bar{\phi}_\mathsf{A}^\top(\boldsymbol{\Lambda}_{AA}-\boldsymbol{\Lambda}_{AB}\boldsymbol{\Lambda}_{BB}^{-1}\boldsymbol{\Lambda}_{BA})\bar{\phi}_\mathsf{A}\right)$$

- この指数部も正規分布の指数部と同じ形をしているので、$P(\phi_A(x))$は平均$\boldsymbol{\mu}_A$、精度行列$\Lambda_{AA} - \Lambda_{AB}\,\Lambda_{BB}^{-1}\,\Lambda_{BA}$（もしくは分散共分散行列$(\Lambda_{AA} - \Lambda_{AB}\,\Lambda_{BB}^{-1}\,\Lambda_{BA})^{-1}$をもつような多次元正規分布となることがわかる

# ところで、分割された行列と、その逆行列には、便利な関係が知られています

- ところで、行列の分割については、以下の等式が知られている

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}$$

  - $\mathbf{M} \equiv (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$ ：シューア補行列と呼ばれる行列

- 分割された行列の逆行列は、各分割部分とその逆行列によって、やはり分割された形で書くことができる

# これを用いると、周辺分布の分散共分散行列は元々のそれの部分行列であることがわかります

- 先の等式に精度行列の分割を入れてみる；
  - $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{\Lambda}_{AA} & \mathbf{\Lambda}_{AA} \\ \mathbf{\Lambda}_{BA} & \mathbf{\Lambda}_{BB} \end{bmatrix}$ として、先の等式に代入
  - すると： $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Lambda}_{AA} & \mathbf{\Lambda}_{AA} \\ \mathbf{\Lambda}_{BA} & \mathbf{\Lambda}_{BB} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{\Lambda}_{AA} - \mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA} & * \\ * & * \end{bmatrix}$

$$\Longleftrightarrow \quad \mathbf{\Sigma} = \begin{bmatrix} (\mathbf{\Lambda}_{AA} - \mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA})^{-1} & * \\ * & * \end{bmatrix}$$

- 分散共分散行列を$\Lambda$と同様の分割： $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{AA} & \mathbf{\Sigma}_{AA} \\ \mathbf{\Sigma}_{BA} & \mathbf{\Sigma}_{BB} \end{bmatrix}$

を考えた時、$P(\phi_{\mathrm{A}}(x))$は、平均$\boldsymbol{\mu}_{\mathrm{A}}$、分散共分散行列$\mathbf{\Sigma}_{\mathrm{AA}}$（つまり、単にもともとの分散共分散行列 $\mathbf{\Sigma}$ の部分行列）をもつ多次元正規分布であるということもできる

# 条件付き分布は、周辺分布の計算で出てきた結果を用いて求めることができます

- つぎに、$\phi_{\mathrm{A}}(x)$ が与えられた時の $\phi_{\mathrm{B}}(x)$ の条件付き確率を求める：

$$P(\phi_{\mathsf{B}}(x)|\phi_{\mathsf{A}}(x)) = \frac{P(\phi(x))}{\phi_{\mathsf{A}}(x)}$$

- 先ほど求めた$P(\phi_{\mathrm{A}}(x))$を使うと、$\mathrm{P}(\phi_{\mathrm{B}}(x)\ |\ \phi_{\mathrm{A}}(x))$は：

$$P(\phi_{\mathsf{B}}(x)|\phi_{\mathsf{A}}(x)) = \frac{\frac{1}{(2\pi)^{D/2}}|\mathbf{\Lambda}|^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\phi(x)-\boldsymbol{\mu})^{\top}\mathbf{\Lambda}(\phi(x)-\boldsymbol{\mu})\right)}{\frac{1}{(2\pi)^{M/2}}|\mathbf{\Lambda}|^{\frac{1}{2}}|\mathbf{\Lambda}_{BB}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\bar{\phi}_{\mathsf{A}}^{\top}(\mathbf{\Lambda}_{AA}-\mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA})\bar{\phi}_{\mathsf{A}}\right)}$$

$$= \frac{1}{(2\pi)^{(D-M)/2}}|\mathbf{\Lambda}_{BB}|^{\frac{1}{2}}$$

$$\exp\left(-\frac{1}{2}(\phi(x)-\boldsymbol{\mu})^{\top}\mathbf{\Lambda}(\phi(x)-\boldsymbol{\mu})+\frac{1}{2}\bar{\phi}_{\mathsf{A}}^{\top}(\mathbf{\Lambda}_{AA}-\mathbf{\Lambda}_{AB}\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA})\bar{\phi}_{\mathsf{A}}\right)$$

- $P(\phi_{\mathrm{A}}(x))$の計算に用いた、指数部分の展開式を使うと、上式の指数部分の第2項が消え：

$$P(\phi_{\mathsf{B}}(x)|\phi_{\mathsf{A}}(x)) =$$

$$\frac{1}{(2\pi)^{(D-M)/2}}|\mathbf{\Lambda}_{BB}|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\bar{\phi}_{\mathsf{B}}+\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)^{\top}\mathbf{\Lambda}_{BB}\left(\bar{\phi}_{\mathsf{B}}+\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)\right)$$

# その精度行列は、もともとの精度行列の部分行列となっていることがわかります

- 求まった形もまた多次元正規分布の形をしている：

$$P(\phi_{\mathsf{B}}(x)|\phi_{\mathsf{A}}(x)) =$$
$$\frac{1}{(2\pi)^{(D-M)/2}}|\mathbf{\Lambda}_{BB}|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\bar{\phi}_{\mathsf{B}}+\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)^{\top}\mathbf{\Lambda}_{BB}\left(\bar{\phi}_{\mathsf{B}}+\mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}\right)\right)$$

- その平均は、$\bar{\phi}_{\mathsf{B}} \equiv \phi_{\mathsf{B}}(x) - \boldsymbol{\mu}_{\mathsf{B}}$ であったことに注意すると、$\boldsymbol{\mu}_{\mathsf{B}} - \mathbf{\Lambda}_{BB}^{-1}\mathbf{\Lambda}_{BA}\bar{\phi}_{\mathsf{A}}$ となることがわかる
- また、その精度行列は $\boldsymbol{\Lambda}_{\mathrm{BB}}$となる
  - 条件付き分布のほうは精度行列で表現したほうがシンプル

  ⇔ 周辺分布は分散共分散行列を使った方がシンプル