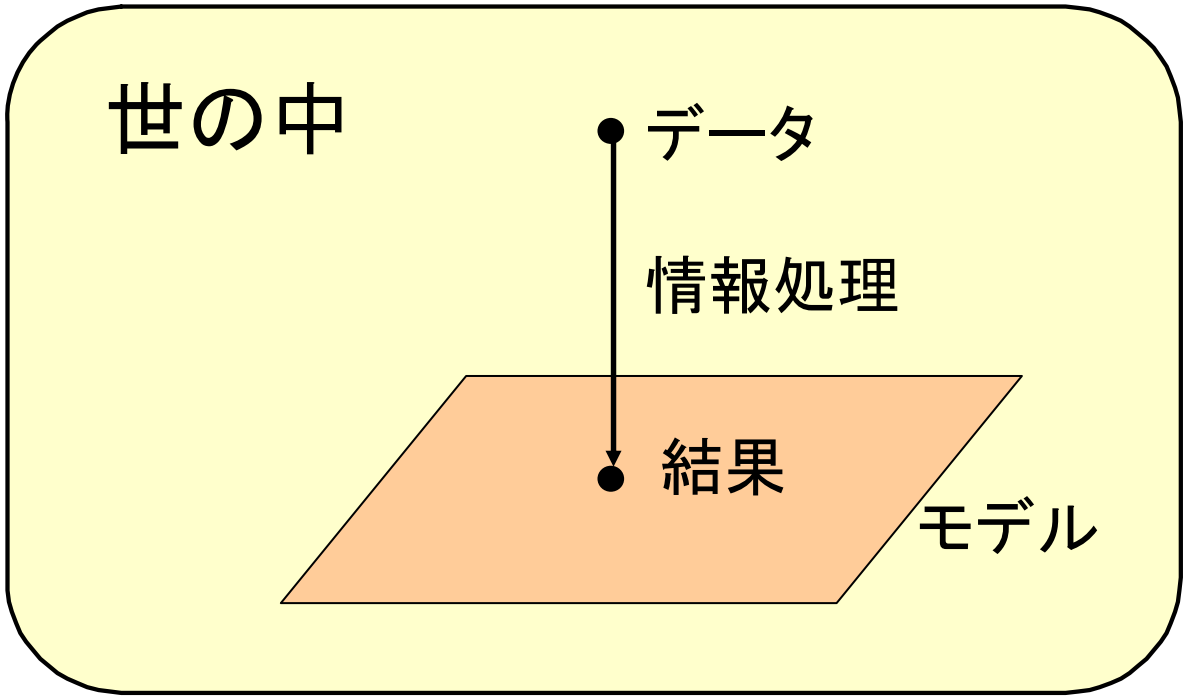
情報幾何入門

赤穂昭太郎
産業技術総合研究所
脳神経情報研究部門

情報幾何

情報処理を幾何的に（図で）理解する

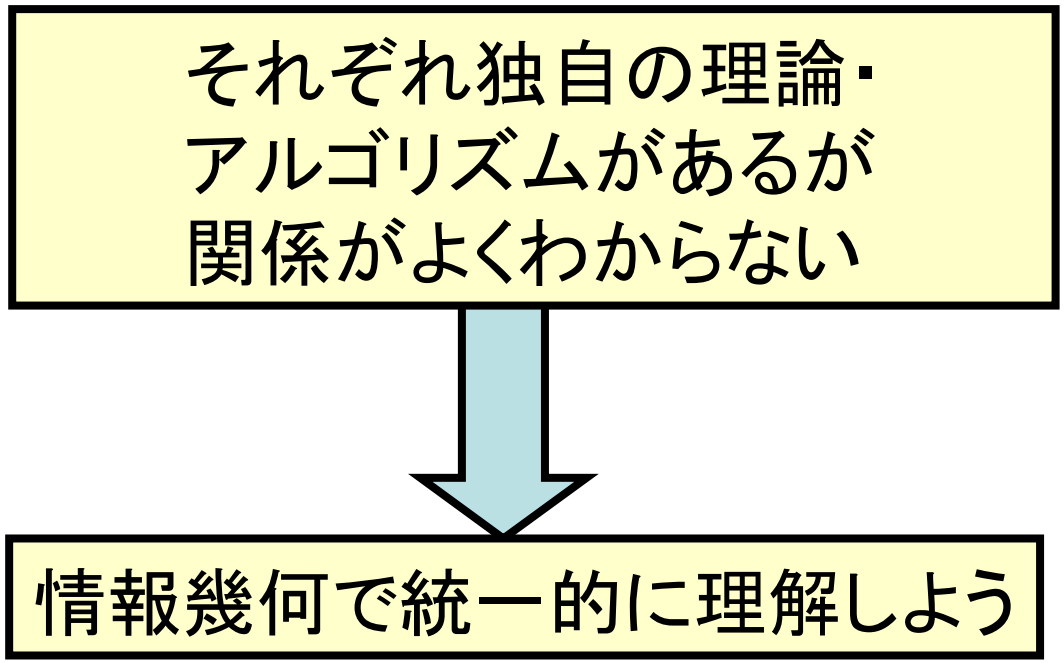


情報幾何から導かれる結論

- 多くのモデルは「平ら」である
- 多くのアルゴリズムは平らなモデルに「まっすぐ」射影を下ろしたものになっている
- ただし，「平ら」「まっすぐ」は普通と違って2種類ある（eとm：双対構造)

共通言語としての情報幾何

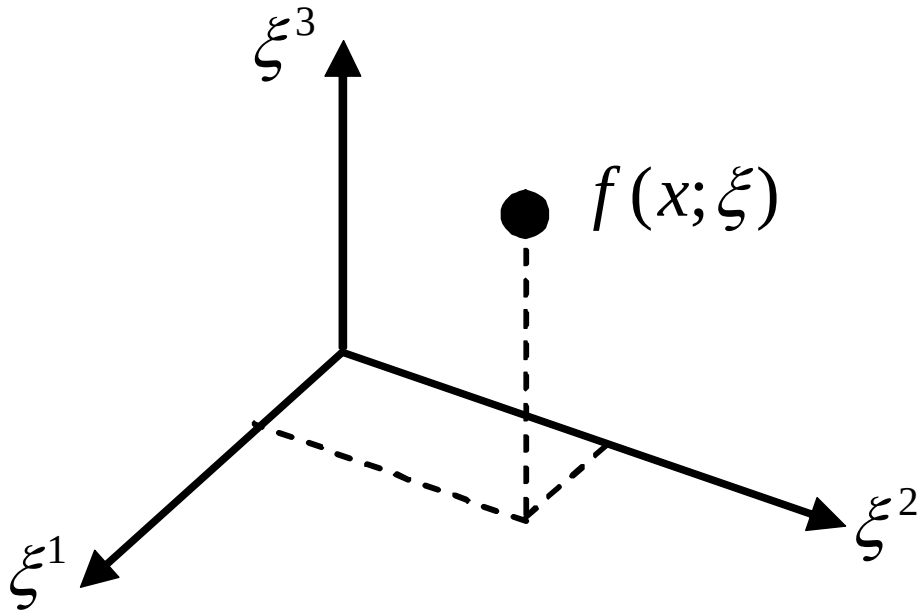
- 確率モデルやその周辺分野
  - 統計学
  - システム制御
  - 符号理論
  - 最適化理論
  - 統計物理



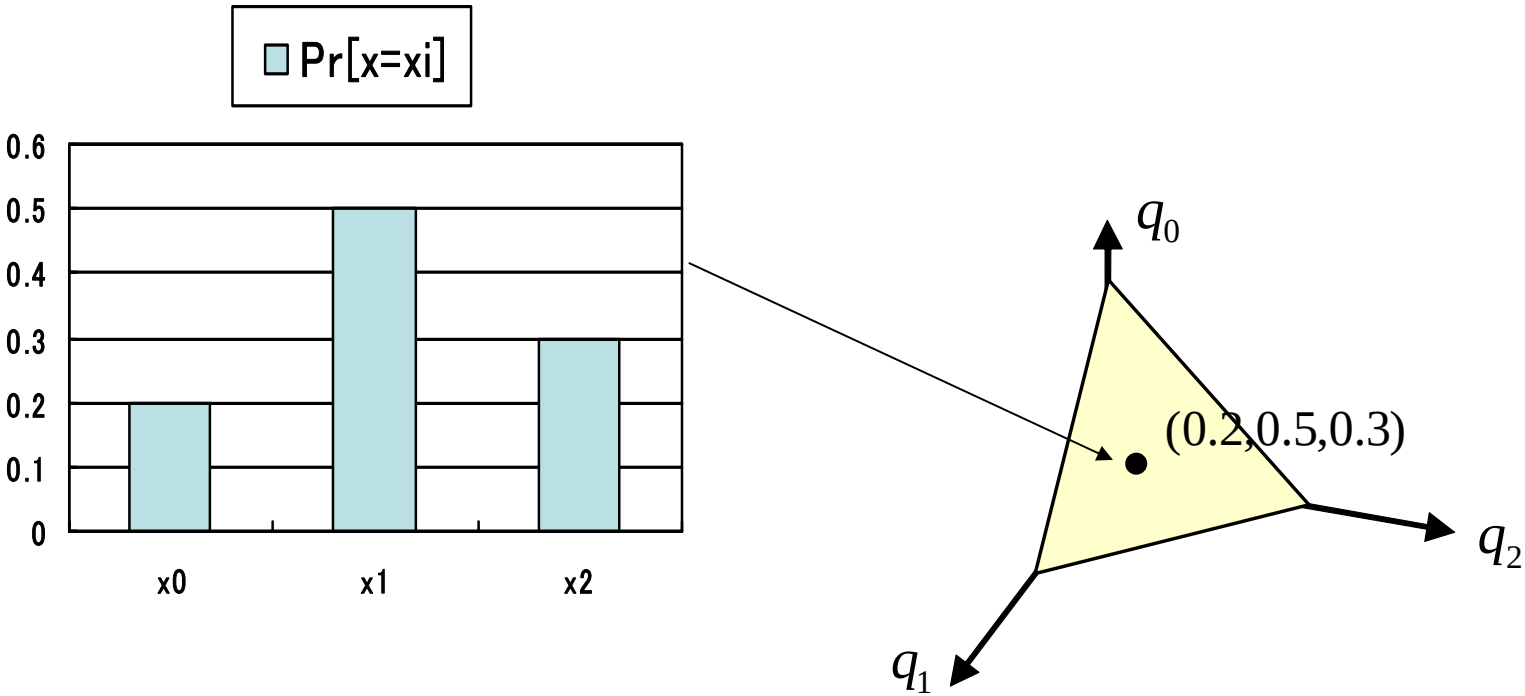
世の中＝確率モデル

- 情報幾何の出発点：
  確率モデル　$f(x;\xi)$　$\xi=(\xi^1,\xi^2,\cdots,\xi^n)$

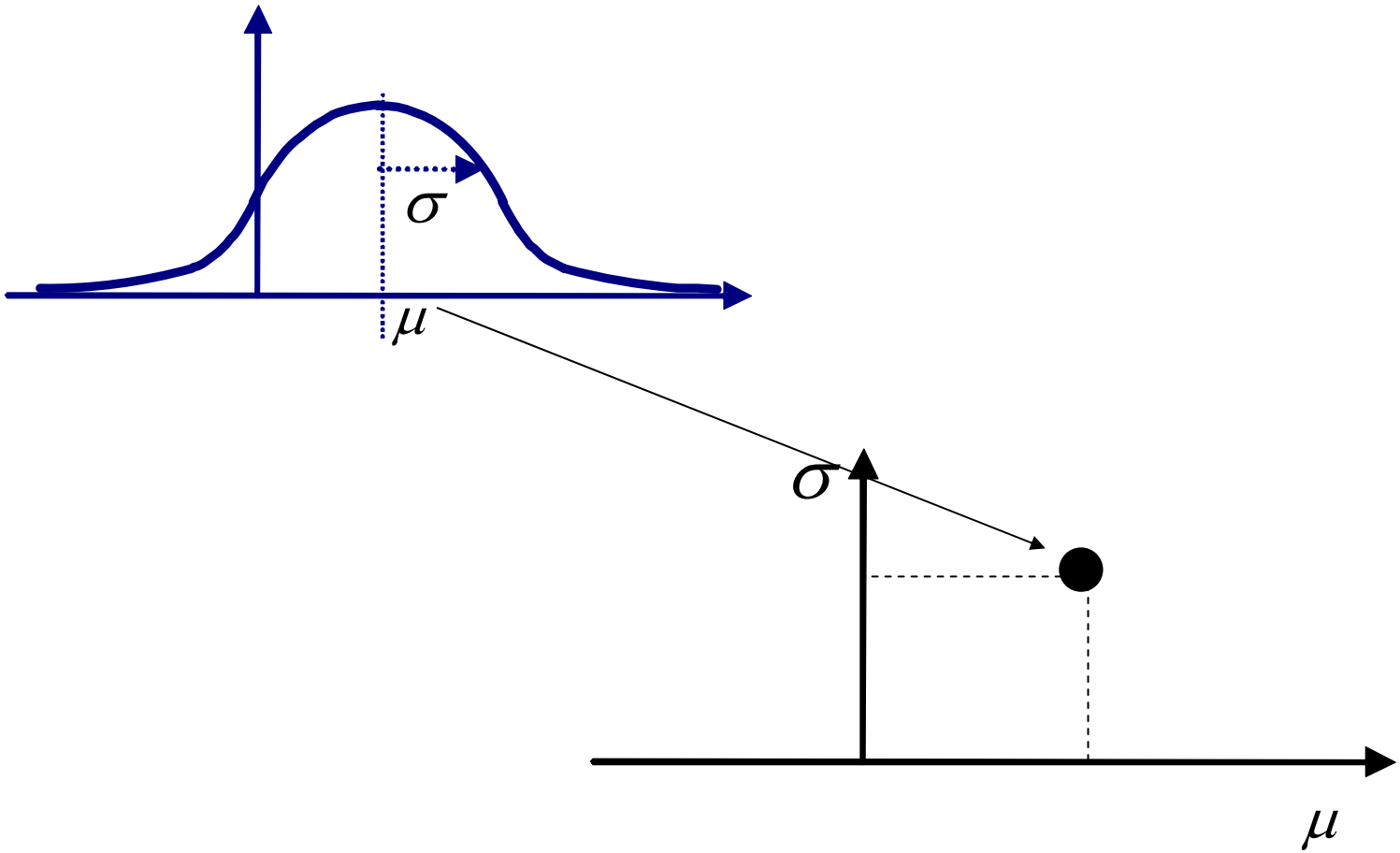
- 座標系



例：離散分布

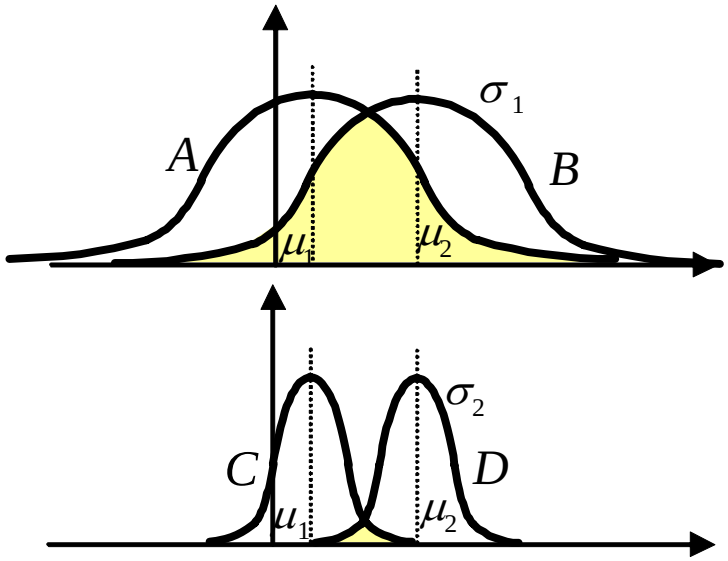


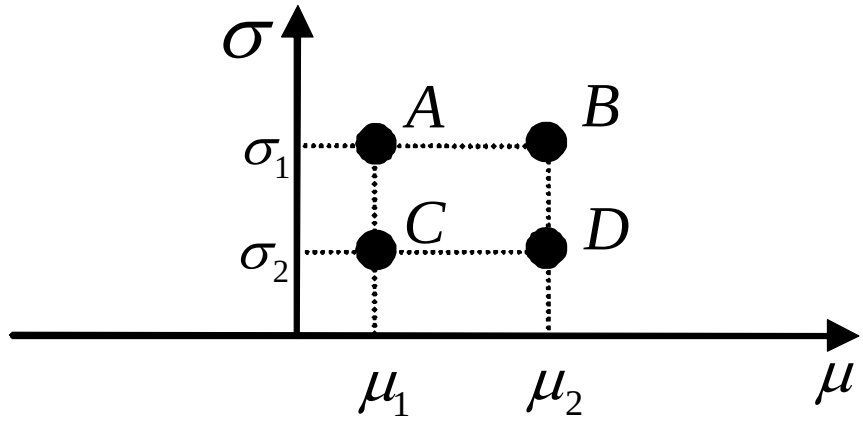
例：正規分布



空間の構造

- ユークリッド空間ではダメ？

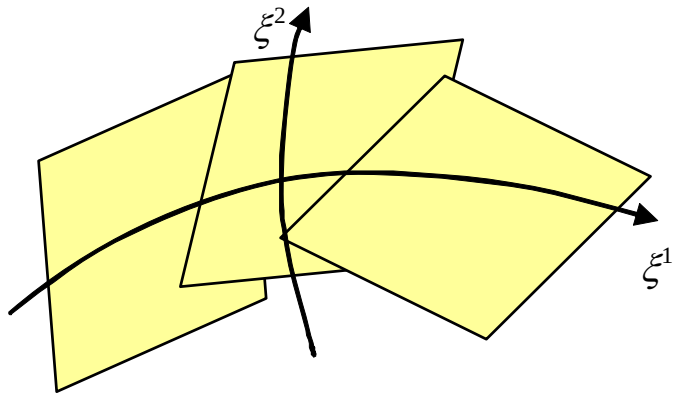




- ユークリッドではA-B と C-D の隔たりが同じになる

空間の構造

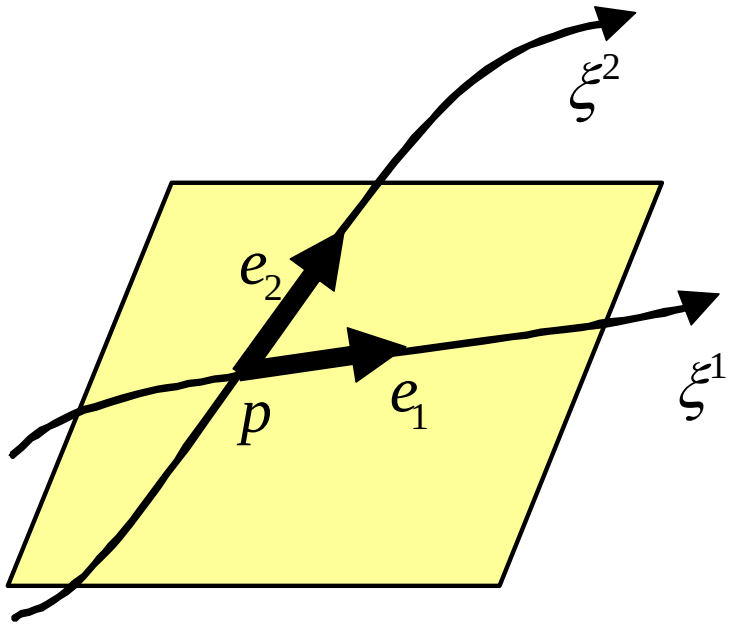
- 空間の構造は何で決まるか？
  - 点の近く: 線形空間（**計量**）
  - 空間全体: 線形空間のつながり方を決める（**接続**）



- 設計方針
  - 統計的に自然なもの
  - パラメータの取り方によらない

点の近くの構造：線形空間

- 線形空間（**接空間**）



- 接空間の構造は
  基底の間の内積で
  決まる（**リーマン計量**）

$$g_{ij} = \left\langle e_i, e_j \right\rangle_\xi$$

情報幾何での計量

- **統計的不変性⇒フィッシャー情報行列**

$$g_{ij}(\xi) = \mathrm{E}_{\xi}[\partial_i \log p(x,\xi) \partial_j \log p(x,\xi)]$$

$$\partial_i = \frac{\partial}{\partial \xi_i}$$

$$\mathrm{E}_{\xi}[f(x)] = \int f(x) p(x;\xi) dx$$

なぜフィッシャー情報量か？

- **クラメール･ラオの不等式**
  N個のサンプルからの ξ の推定量 $\hat{\xi}$ の分散の下限

$$\mathrm{Var}[\hat{\xi}] \geq \frac{1}{N} G^{-1}(\xi)$$

- G^{-1}が ξ のまわりでの散らばり具合を表す

  ⇔ G^{-1}が大きいところはきめが粗い

例：正規分布

$$p(x;\mu,\sigma)=\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}-\frac{1}{2}\log 2\pi\sigma^2\right)$$

$$G=\frac{1}{\sigma^2}\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix}$$

- $d\mu, d\sigma$ だけ微小に動かしたときの変化は $(d\mu^2+2d\sigma^2)/\sigma^2$
  ⇒分散の小さいところは少し動かしただけで大きくずれる

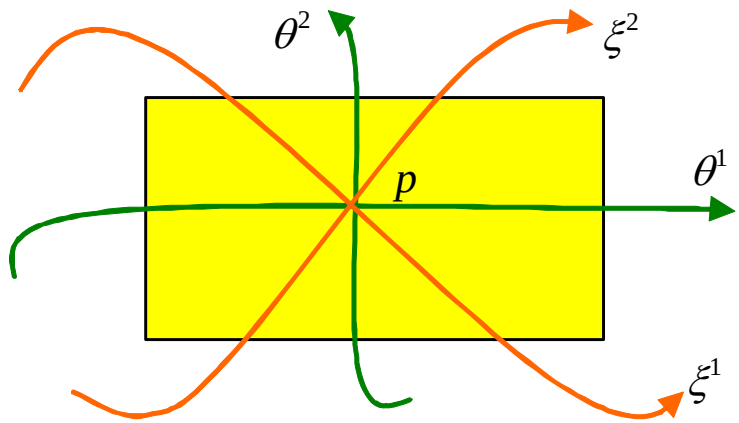
計量と座標変換

- 計量は(一般に非線形な)座標変換に対して線形に変換される(テンソル)

$$\xi = \left(\xi^i\right) \mapsto \theta = \left(\theta^a\right)$$

$$g_{ij} = \sum_{a,b} J_i^a J_j^b g_{ab}$$

$$J_i^a = \frac{\partial \theta^a}{\partial \xi^i}$$

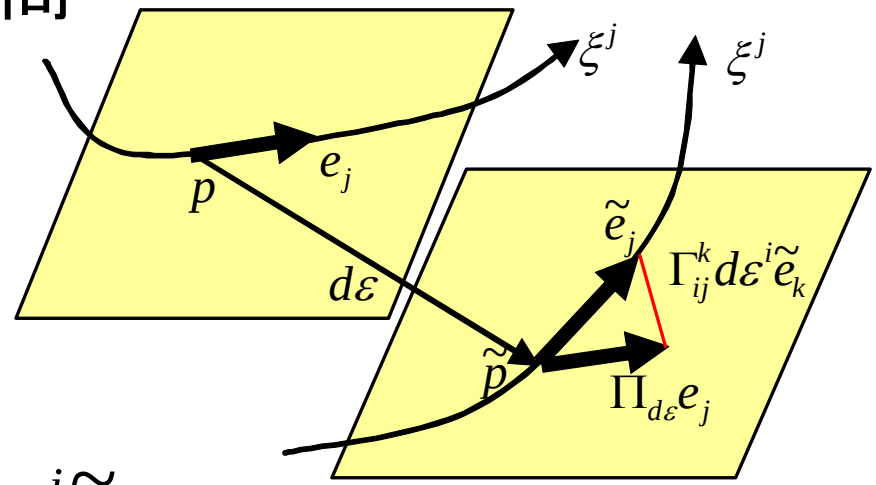


ユークリッド空間をつなぐ

- 各点ごとにバラバラの接空間

$$\xi(\tilde{p}) = \xi(p) + d\varepsilon$$

⇒接空間をつなぐ（接続）



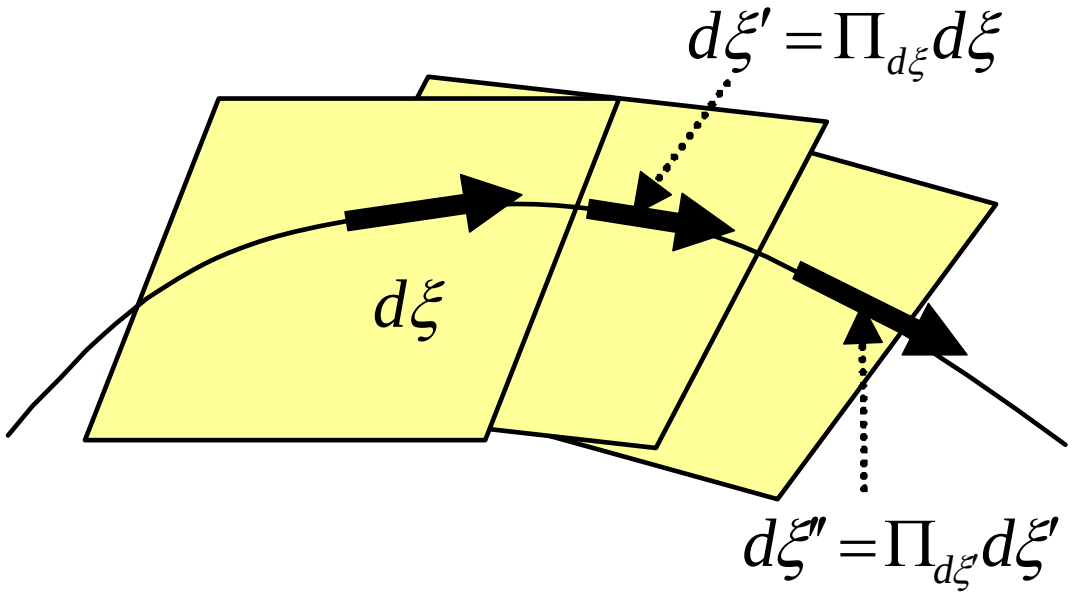
- 接ベクトルe_jの平行移動

$$\Pi_{d\varepsilon}\left[e_j\right] = \tilde{e}_j - \sum_{i,k} \Gamma_{ij}^k d\varepsilon^i \tilde{e}_k$$

- Γ_{ij}^kを（アファイン）接続係数と呼ぶ

測地線：まっすぐな線

- ある接ベクトルの方向 $d\xi$ の自分自身への平行移動 $\Pi_{d\xi}[d\xi]$ をつなげたものを**測地線**という（直線の概念の一般化）



接続をどう決めるか？

- 二つの接ベクトルを平行移動したとき，
  普通（物理等）はその間の内積を保存したい

$$\left\langle \Pi_{d\varepsilon}\left[d\xi_1\right], \Pi_{d\varepsilon}\left[d\xi_2\right] \right\rangle = \left\langle d\xi_1, d\xi_2 \right\rangle$$

- これを満たす接続は計量から一意的に決まってしまう⇒レビ・チビタ接続
- ところが情報幾何ではそれ以外の接続も考える

α 接続

- 統計的な不変性⇒パラメータαをもつ接続係数に限られる

$$\Gamma_{ij,k}^{(\alpha)}(\xi) = \mathrm{E}_{\xi}\left[\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right]$$

$$\partial_i l = \frac{\partial}{\partial \xi_i}\log p(x;\xi) \qquad \Gamma_{ij,k} = \sum_h \Gamma_{ij}^h g_{hk}$$

- 特にα＝0のときがレビ・チビタ接続
- 情報幾何ではα＝±1のときが最重要！

平坦な空間

- 接続はテンソルではない（座標系に依存）
- 逆に言えば，うまく座標系を取れば，$\Gamma=0$にできる(まっすぐな空間)
- このような座標系がもし存在するとき αアファイン座標系といい，その座標系についてα平坦であるという.
- 平坦な座標系の測地線（α測地線）はαアファイン座標系での直線になっている.

$$\xi = (1-t)\xi_0 + t\xi_1$$

重要な分布族

- $\alpha = \pm 1$ は特別な意味がある:

確率分布の分布族で，α 平坦になるのは
「指数分布族(exponential family)」と
「混合分布族(mixture family)」の
二つだけで，それぞれ $\alpha = \pm 1$ に対応する

指数分布族

- 情報幾何で最も基本的な分布族

$$p(x;\xi) = \exp\left(\sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta) + C(x) \right)$$

- 指数分布族は θ をアファイン座標系として1-平坦
- 指数分布族は特別なので1-平坦や1-接続のことをe-平坦とかe-接続という(e=exponential)

混合分布族

- 確率分布の線形和

$$p(x;\xi) = \sum_{i=1}^{n} \theta^i F_i(x) + \theta^0 F_0(x)$$

$$\theta^0 = 1 - \sum_{i=1}^{n} \theta^i$$

- パラメータθをアファイン座標系として
  －1平坦
- 混合分布族は特別なので－1平坦，－1接続のことをm平坦，m接続という(m:mixture)

離散分布は混合かつ指数

- 混合分布族としては

$$p(x;\xi) = \sum_{i=1}^{n} q_i \delta(x-i) + q_0 \delta(x)$$

- 指数分布族としては

$$p(x;\xi) = \exp\left(\sum_{i=1}^{n} r_i \delta(x-i) - \psi(r) \right)$$

$$r_i = \log q_i - \log q_0 \qquad \psi(r) = -\log q_0$$

正規分布は指数分布族

$$p(x;\mu,\sigma)=\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}-\frac{1}{2}\log 2\pi\sigma^2\right)$$

$$p(x;\xi)=\exp\left(\sum_{i=1}^{n}\theta^i F_i(x)-\psi(\theta)+C(x)\right)$$

$$F_1(x)=x \qquad \theta^1=\mu/\sigma^2$$

$$F_2(x)=x^2 \qquad \theta^2=-1/(2\sigma^2)$$

$$\psi(\theta)=\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\log 2\pi\sigma^2 \qquad C(x)=0$$

双対平坦と双対座標

- 実は α 平坦なら，別の座標系が存在して
  $-\alpha$ 平坦 になる
- α 平坦な座標系: θ ，$-\alpha$ 平坦な座標系: η
- ルジャンドル変換:ポテンシャル関数 ψ , φ

$$\psi(\theta)+\varphi(\eta)-\sum_i \theta^i \eta_i = 0$$

$$\frac{\partial \varphi(\eta)}{\partial \eta} = \theta \qquad \frac{\partial \psi(\theta)}{\partial \theta} = \eta$$

双対性

- θ に対する計量: g_{ij}　η に対する計量: g^{ij}

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij} \qquad \frac{\partial \theta^i}{\partial \eta_j} = g^{ij}$$

- 計量が座標変換のヤコビ行列になっている
- θ 座標での基底: e_i　η 座標での基底: e^j

双対直交: $\left\langle e_i, e^j \right\rangle = \delta_i^j$

指数分布族の場合

- θ 座標系は1平坦

$$p(x;\xi) = \exp\left(\sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta) + C(x) \right)$$

- 双対座標は $\eta_i = \mathrm{E}_\theta\left[F_i(x)\right]$
- ポテンシャルは ψ そのもの
- 混合分布族も双対平坦だが双対座標が単純な形で書けないので，結局指数分布族が唯一重要な分布族

離散分布の場合

- e座標系 $p(x;\xi)=\exp\left(\sum_{i=1}^{n} r_i \delta(x-i)-\psi(r)\right)$

$$\theta^i = r_i = \log q_i - \log q_0$$

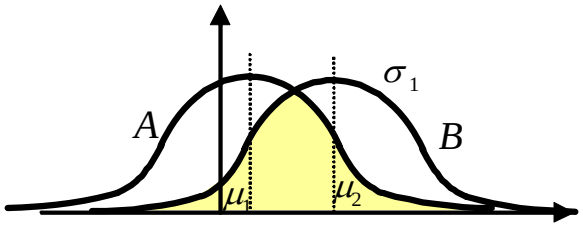
確率値の対数の線形空間

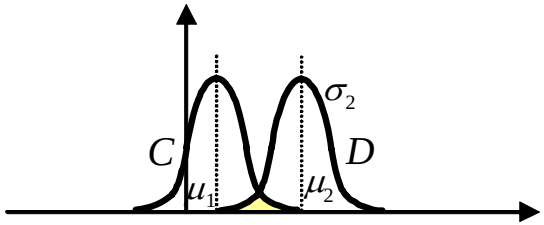
- m座標系

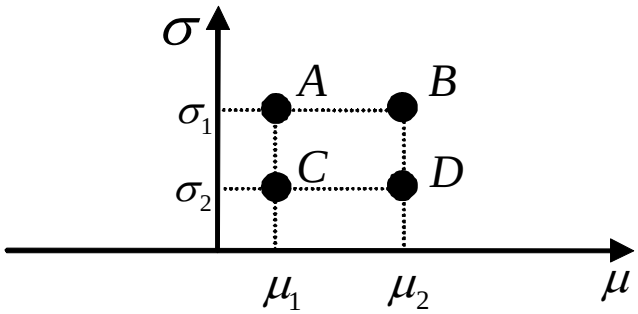
$$\eta_i = E_\theta[\delta(x-i)] = q_i$$

確率値の線形空間

例：正規分布

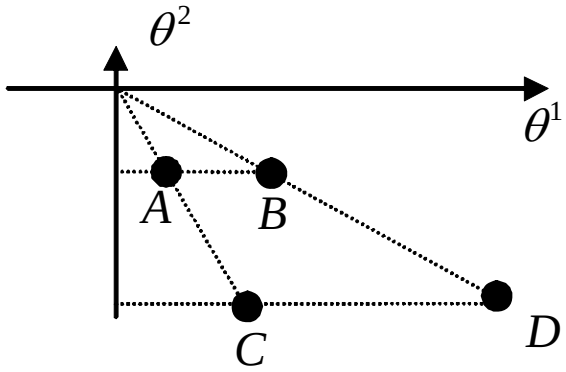






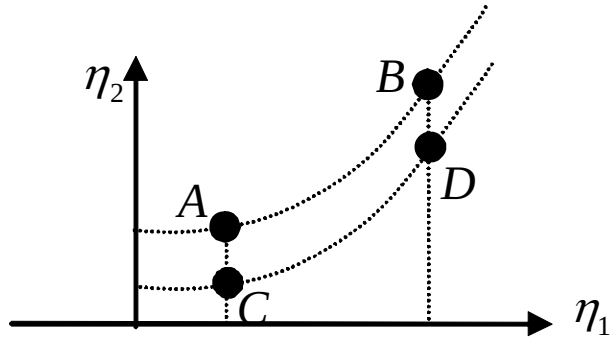
$$\theta^1 = \mu / \sigma^2$$

$$\theta^2 = -1 / (2\sigma^2)$$



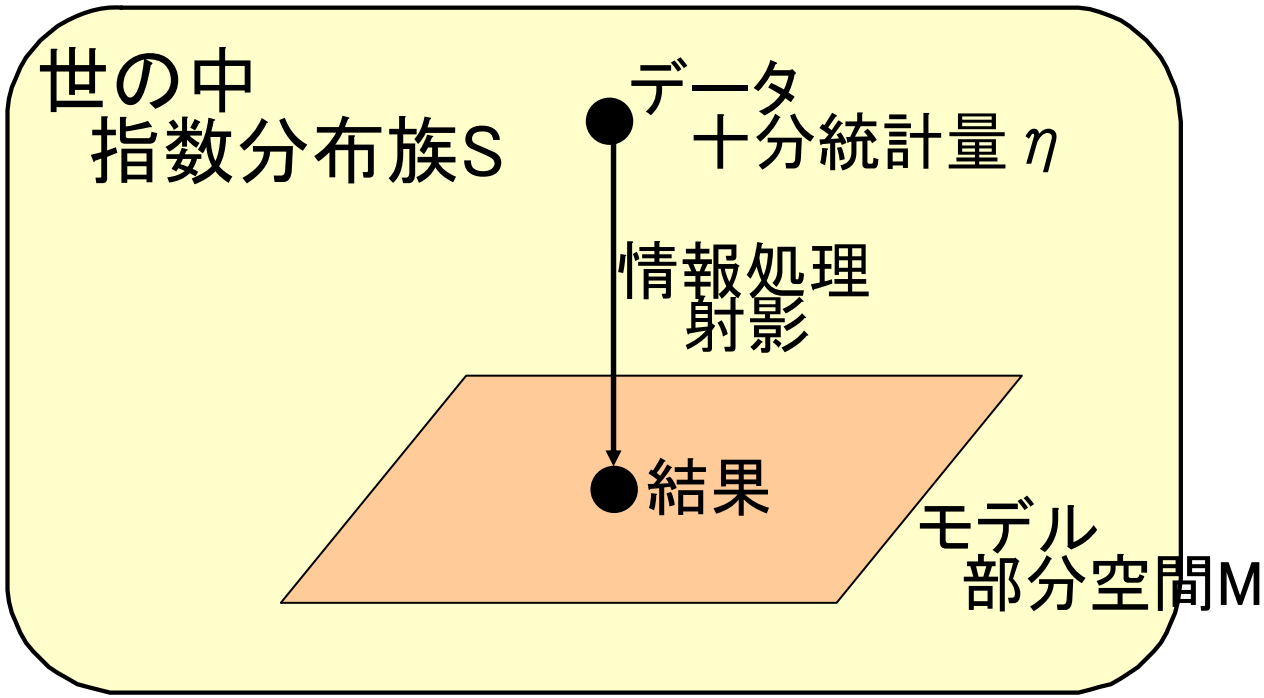
$$\eta_1 = \mathrm{E}_\theta[x] = \mu$$

$$\eta_2 = \mathrm{E}_\theta[x^2] = \mu^2 + \sigma^2$$



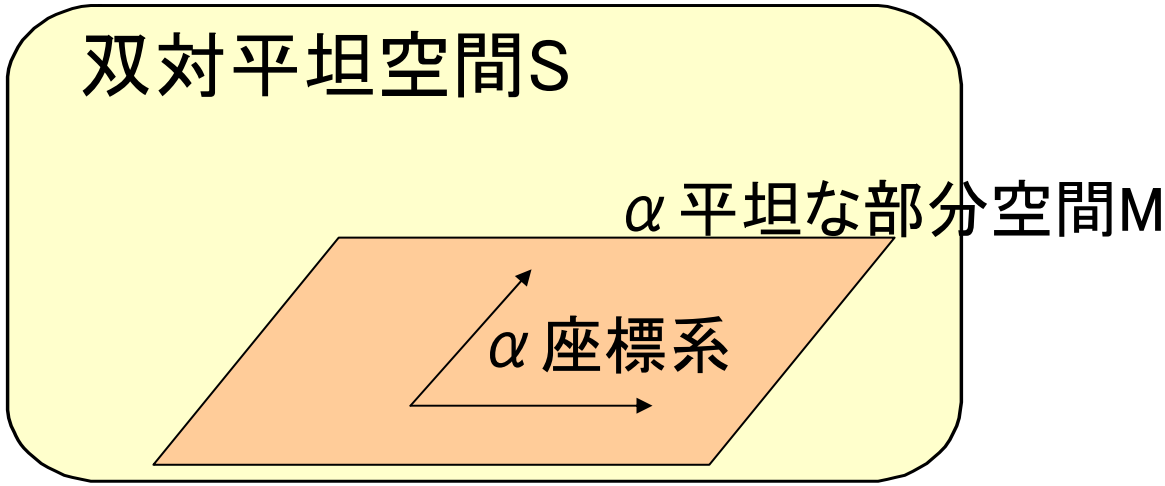
部分空間と射影

- 情報幾何的世界観



平坦な部分空間

- α平坦な線形部分空間：双対平坦な空間Sのα座標系での線形部分空間



- 注意：α平坦な部分空間は一α平坦な部分空間とは限らない c.f. S自身はどちらも平坦

ダイバージェンス

- 射影を導入する前に. . .
- αダイバージェンス

$$D^{(\alpha)}(p\,\|\,q)=\psi(\theta(p))+\varphi(\eta(q))-\sum_i \theta^i(p)\eta_i(q)$$

c.f. ルジャンドル変換 $\psi(\theta)+\varphi(\eta)-\sum_i \theta^i\eta_i=0$

- 対称律以外は距離の性質を満たす
- p≒q なら距離に一致する
- 双対性 $D^{(\alpha)}(p\,\|\,q)=D^{(-\alpha)}(q\,\|\,p)$

指数分布族の場合

- $\alpha = 1$（e接続）でのダイバージェンスは
  カルバックダイバージェンスに一致する

$$KL(f \| g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

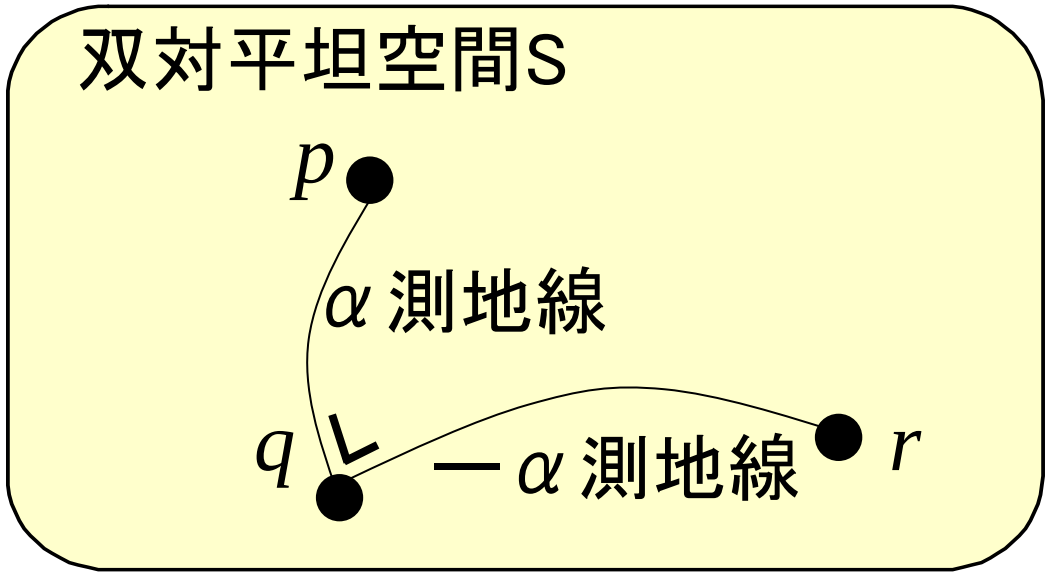
- $\alpha = -1$（m接続）でのダイバージェンスは
  $KL(g \| f)$

距離の分解

- ユークリッド空間で部分空間への射影を取るのがなぜ簡単か？
- ある点から部分空間への距離が直交成分と水平成分に簡単に分解できるから（ピタゴラスの定理）

$$(x-y)^2 = (x-y^{\perp})^2 + (y-y^{\perp})^2$$

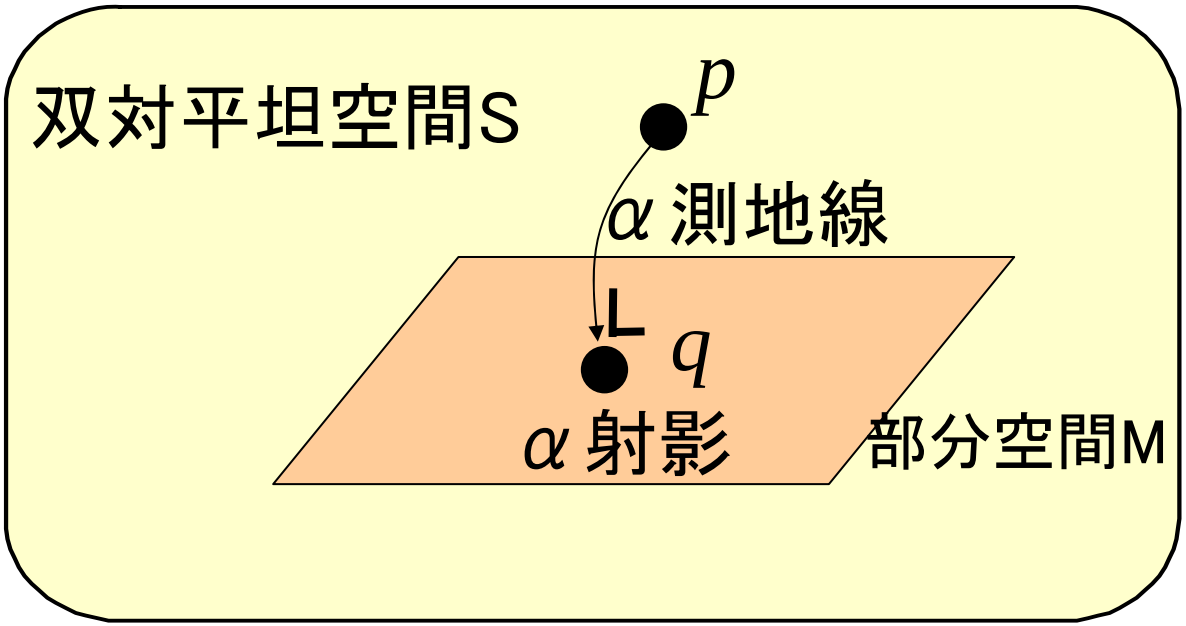
拡張ピタゴラスの定理



$$D^{(\alpha)}(p \| r) = D^{(\alpha)}(p \| q) + D^{(\alpha)}(q \| r)$$

射影定理

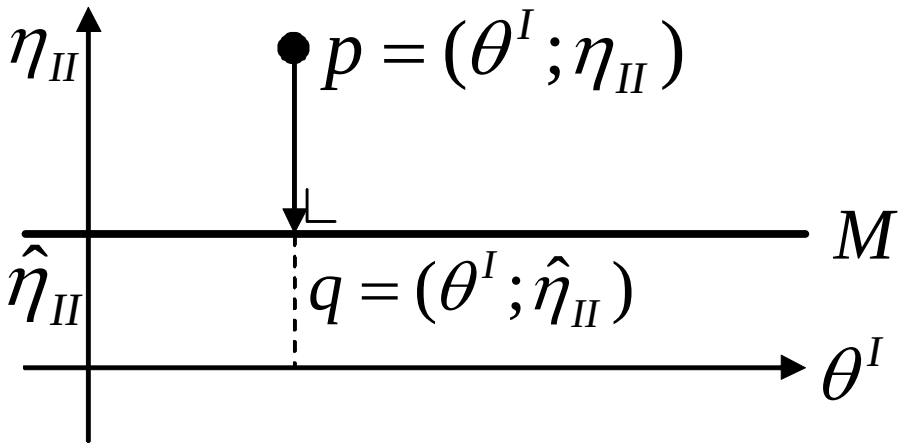
- α 測地線で引いた直交射影は
  α ダイバージェンス $D^{(\alpha)}(p\|q)$ の停留点



- 特にMが一 α 平坦なら $\min_q D^{(\alpha)}(p\|q)$

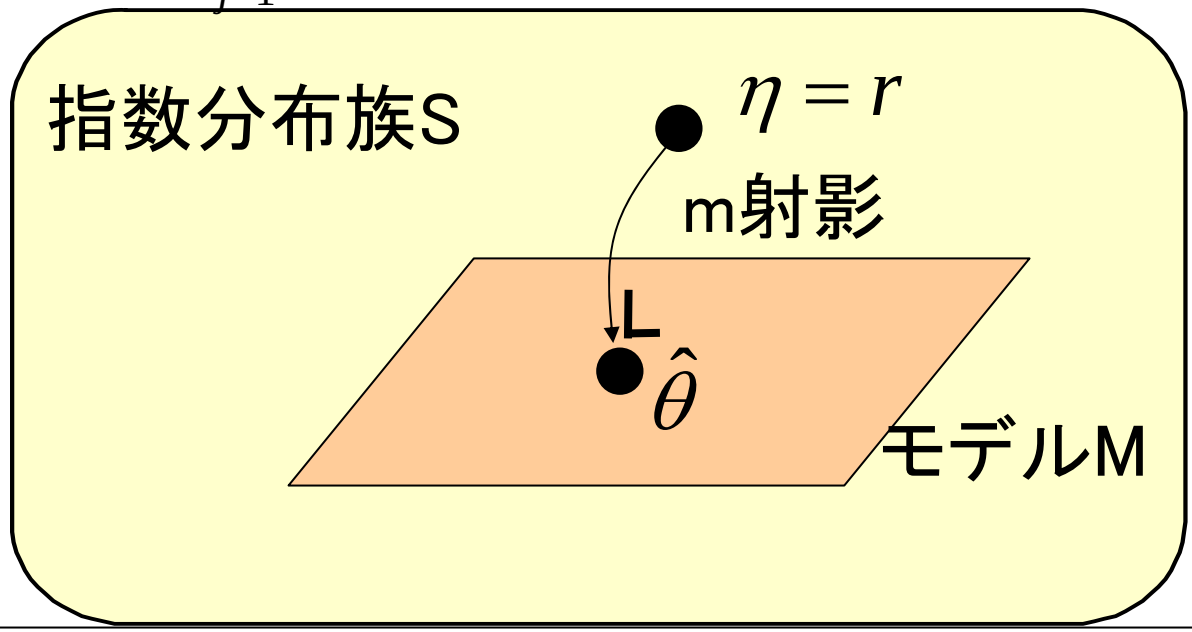
混合座標系：全部まっすぐに見える

- α 射影と $-\alpha$ 部分空間の組み合わせが一番単純
  ←双対性から $\left\langle e_i , e^j \right\rangle = \delta_i^{\ j}$
- Mの中と外とで α 座標系と $-\alpha$ 座標系を分けて使えばまっすぐな図が描け，射影も陽に表現できる



統計的推定

- データは空間のどの点に配置するか？
- $\eta_i = \mathrm{E}_\theta\left[F_i(x)\right]$　なので，N個のデータの十分統計量 $r_i = \frac{1}{N}\sum_{j=1}^{N} F_i(x^{(j)})$ を η 座標とすればよい



統計的推定（つづき）

- 最尤推定 $\max_{\theta} p(x^{(1)} \cdots, x^{(N)}; \theta)$

$$\Leftrightarrow \max_{\theta \in M} \sum_{j=1}^{N} \log p(x^{(j)}; \theta)$$

- 最尤推定はm射影と等価

$$KL(q(x) \| p(x;\theta)) = \int q(x) \log \frac{q(x)}{p(x;\theta)} dx \to \min_{\theta \in M}$$

- モデルが平らなときは推定が易しい.
  推定の質についてはモデルの曲がり具合（曲率）に関係⇒統計的漸近理論

線形システム

- 線形システム　$\varepsilon(t) \approx N(0,1)$ → $H(z)$ → $x(t)$

$$x(t) = \sum_{i=0}^{\infty} h_i \varepsilon(t-i) = H(z)\varepsilon(t)$$

伝達関数

$$H(z) = \sum_{i=0}^{\infty} h_i z^i$$

パワースペクトラム

$$S(\omega) = | H(e^{i\omega}) |^2$$

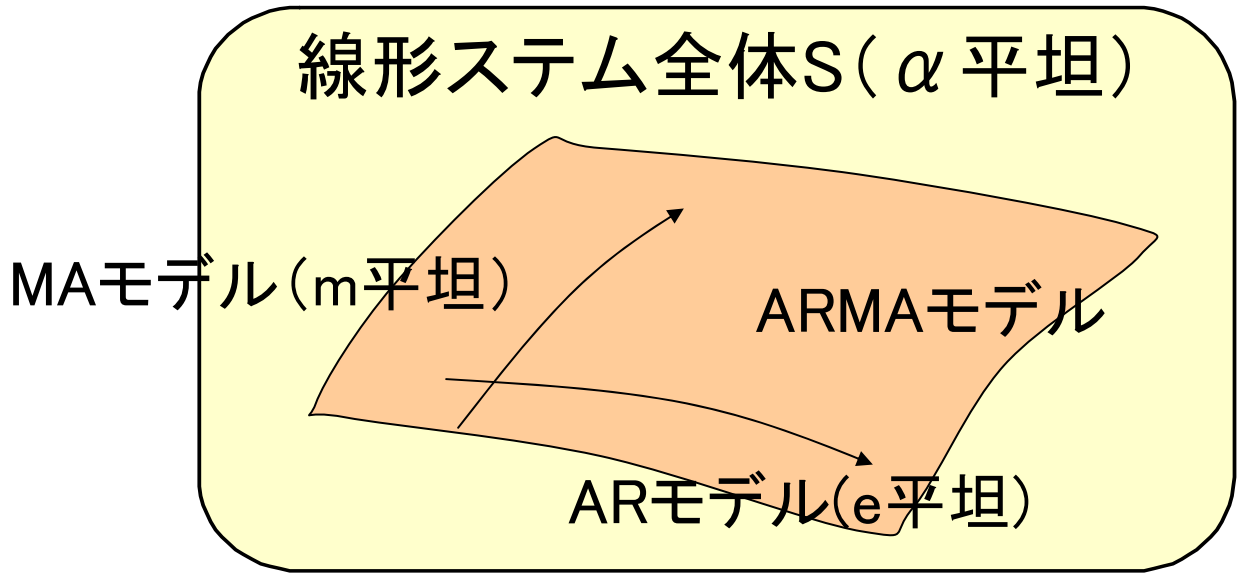
- システムの例：ARモデル，MAモデル，ARMAモデルなど
- 最小位相推移→*H*と*S*が1対1に対応

線形システム（つづき）

- 確率モデル：信号$x(t)$の周波数成分$X(\omega)$

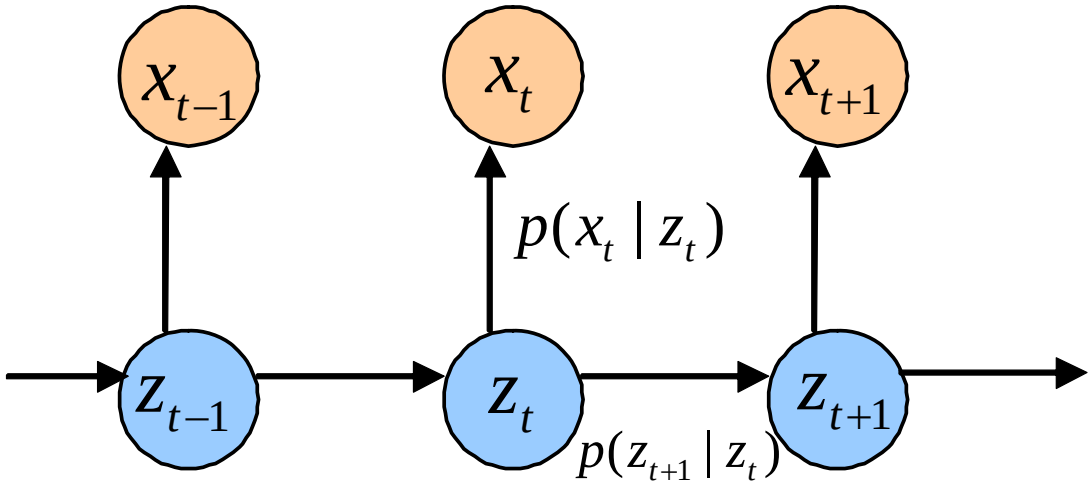
$$p(X;S)=\exp\left(-\frac{1}{2}\int\frac{|X(\omega)|^2}{S(\omega)}-\psi(S)\right)$$

- 実はすべてのαについてα平坦になる



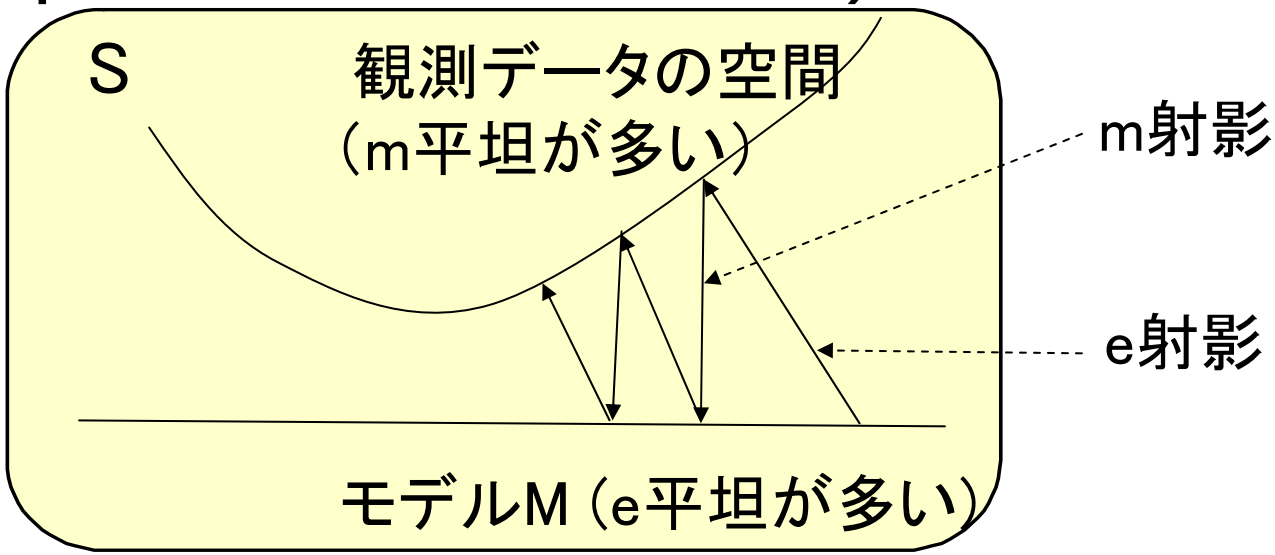
潜在変数モデル

- x だけが観測される　$p(x, z; \xi)$
  例: 隠れマルコフモデル(HMM)



em アルゴリズム

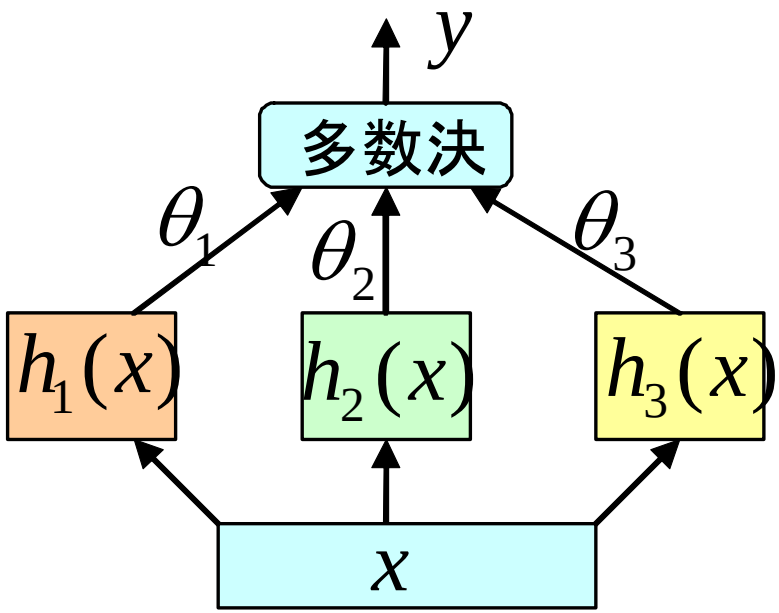
- em (exponential and mixture)



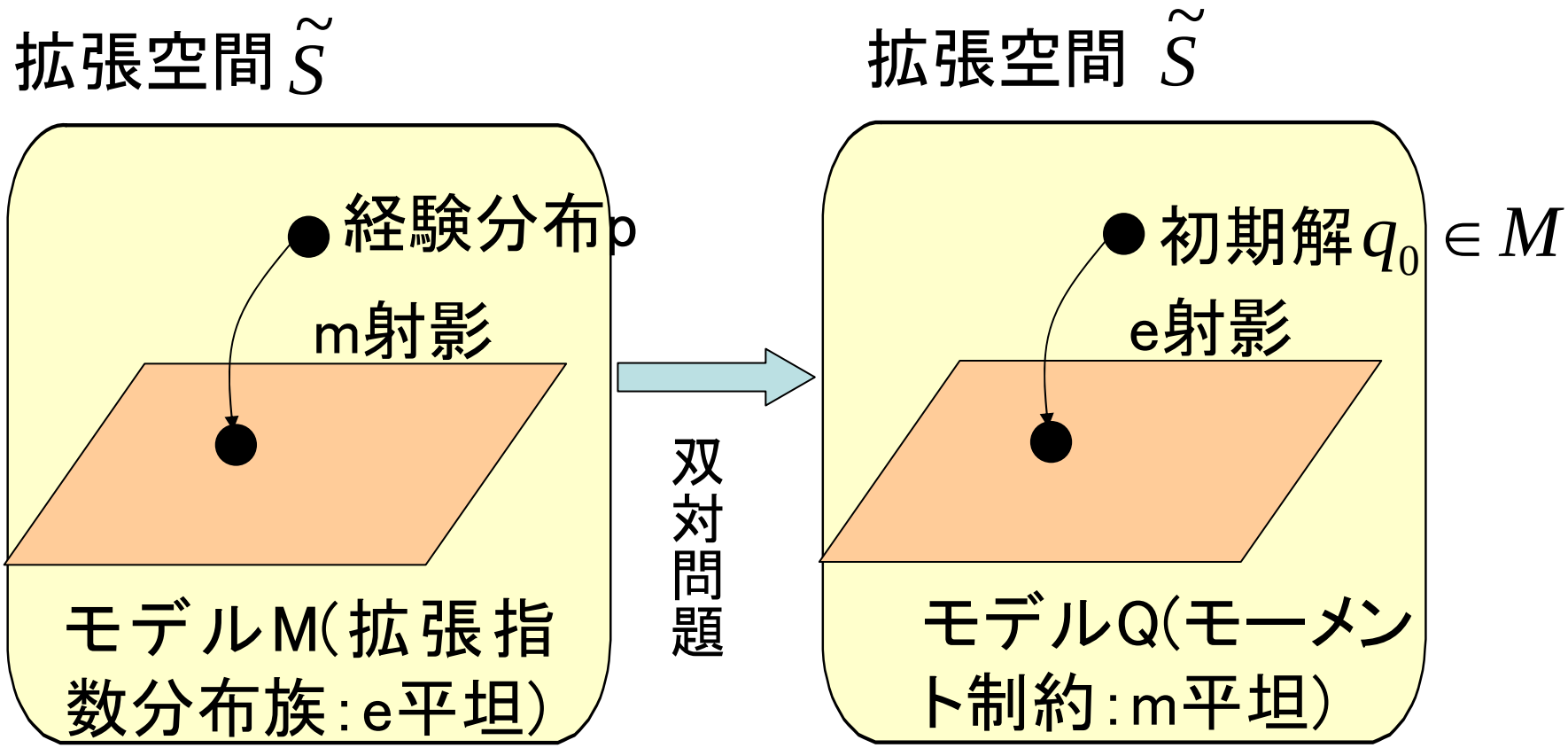
- 実はこれがEMアルゴリズム(Expectation-Maximization/Baum-Welch) とほぼ等価

集団学習

- 三人寄れば文殊の知恵？
- バギング・ブースティング

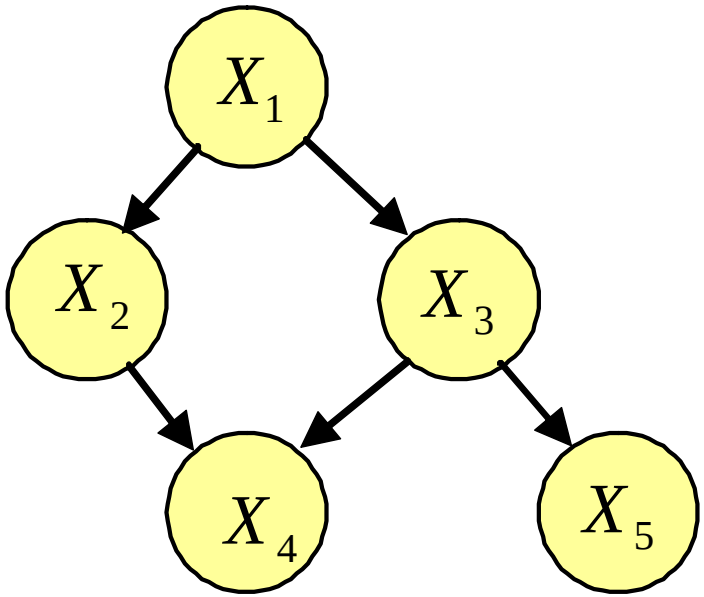


集団学習(つづき)


拡張空間 $\tilde{S}$
経験分布p
m射影
モデルM(拡張指数分布族:e平坦)
双対問題
拡張空間 $\tilde{S}$
初期解 $q_0 \in M$
e射影
モデルQ(モーメント制約:m平坦)

グラフィカルモデルとベイズ推定

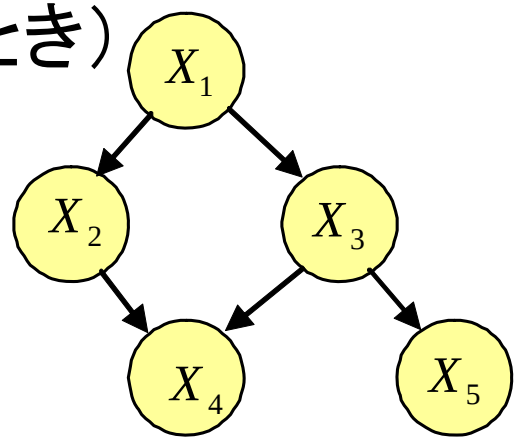
- 変数間の依存関係をグラフであらわす
- HMM, カルマンフィルタもその一種



$$p(X) = p(X_1)$$
$$p(X_2 \mid X_1)p(X_3 \mid X_1)$$
$$p(X_4 \mid X_2, X_3)p(X_5 \mid X_3)$$

ベイズ推定

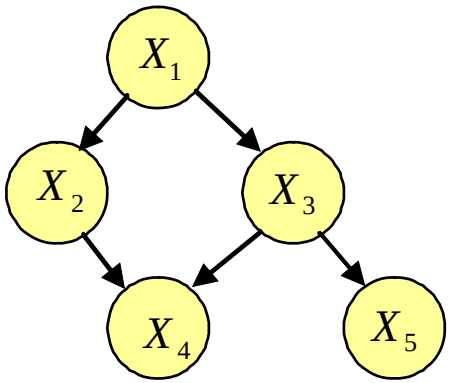
- 一部が観測されたときに残りの変数を推定
  事後分布 $$p(X_1, X_2, X_3 \mid X_4, X_5) = \frac{p(X)}{p(X_4, X_5)} = \frac{p(X)}{\sum_{X_1, X_2, X_3} p(X)}$$
- ノード数が増えると総和計算
  （or 積分）が大変！（特に木でないとき）
- ⇒近似計算
  （平均場近似・変分ベイズ）
  （マルコフ連鎖モンテカルロ・
  パーティクルフィルタ）

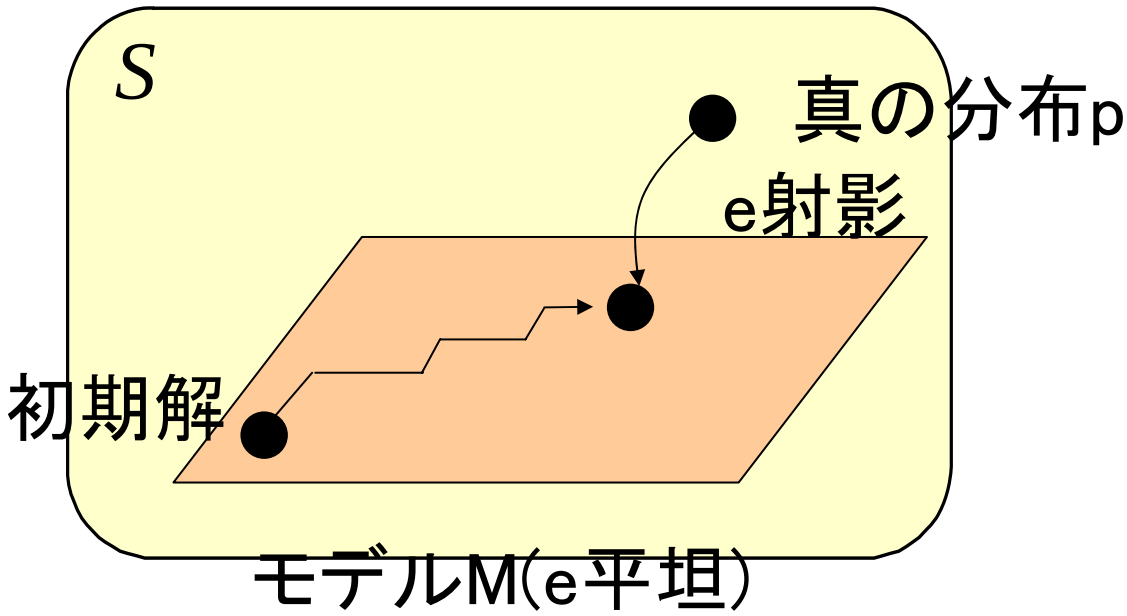


平均場近似・変分ベイズ法

$$p(X_1, X_2, X_3 \mid X_4, X_5) \cong q_1(X_1) q_2(X_2) q_3(X_3)$$ モデルM(e平坦)

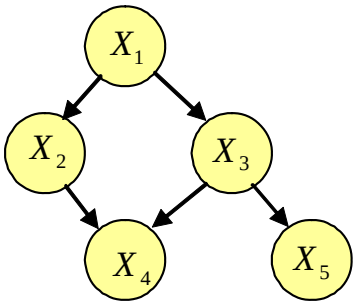
$$\min KL\left[q_1(X_1) q_2(X_2) q_3(X_3) \,\|\, p(X_1, X_2, X_3 \mid X_4, X_5)\right]$$ e射影





マルコフ連鎖モンテカルロ

- 乱数発生により事後分布からのサンプルを生成する
- ギブスサンプラー



$$p(X_1^{(t+1)} \mid X_2^{(t)}, X_3^{(t)}; X_4, X_5)$$

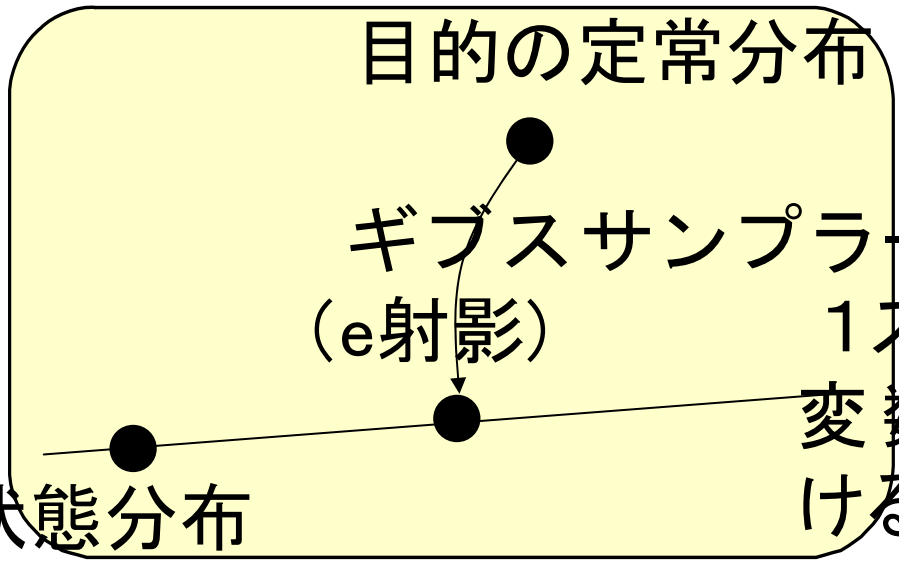
$$p(X_2^{(t+1)} \mid X_3^{(t)}, X_1^{(t+1)}; X_4, X_5)$$

$$p(X_3^{(t+1)} \mid X_1^{(t+1)}, X_2^{(t+1)}; X_4, X_5)$$

- どのような初期値から始めても，
  $p(X_1, X_2, X_3 \mid X_4, X_5)$ に分布収束する

ギブスサンプラーの幾何

- 1ステップに一つの変数を更新するマルコフ連鎖モンテカルロを考える.



さらなる発展

- 有限次元のパラメータ空間から無限次元の空間の幾何へ（セミパラメトリック幾何）
- 特異点の問題（ニューラルネットなどの階層的なモデル：代数幾何の高みへ）
- 新たな情報処理へ．．．

参考文献

- 赤穂：情報幾何と機械学習
（「計測と制御」2005年5月号）
- 甘利：情報幾何とその応用
（「システム・制御・情報」連載
2004年6月～）
- 公文：推定と検定への幾何学的アプローチ，
（「統計科学のフロンティア 2
統計学の基礎II」，岩波書店）