スパースガウス過程回帰

作成者 Onoue Keisuke
作成日 2022/05/14
更新日 2023/05/18

目次

0. 導入

ガウス過程とは、平均関数 $m(\mathbf{x})$と共分散関数 $k(\mathbf{x}, \mathbf{x'})$ によって定義される確率過程で、あらゆる n 個の入力点 $\mathbf{x}_n$ の関数 f による出力値 $f(\mathbf{x}_n)$ の分布が、平均 $m(\mathbf{x})$と $K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'})$ を要素とする分散共分散行列 $\mathbf{K}$ の多次元ガウス分布に従うものというふうに定義されます。

直感的な説明としては、「サイコロ」を振ると 1 ~ 6 の自然数が出てくる箱と表現されるのに対して、「ガウス過程」は振ると関数 $f()$ が出てくる箱のようなものだと言われます。

このガウス過程を、分析したいデータの背後にある構造だと仮定して行う回帰の手法がガウス過程回帰です。

ガウス過程回帰の魅力は様々ですが、非線形な関係性をモデル化できる、つまりモデルとしての表現能力がとても高いことや、予測値に対しての自信の度合いを出力できることなどが挙げられます。

対して、計算量がデータ点数の３乗に比例するために、データ点数が大きくなると現実的な時間内で計算を終えるのが難しくなるといった問題もあります。

このガウス過程回帰の最大の問題点である多大な計算コストに対応するための手段の１つとして、スパース近似（補助変数法）などが提案されてきました。

この記事は、直感的にわかりやすい説明というよりは、「ガウス過程と機械学習（講談社）」を読んでいて私が躓いた箇所、式変形や論理展開などを補足する形で書いたものです。

以下を導出することをこの記事のゴールとします。

- スパースガウス過程回帰モデルの予測分布

$$p(\mathbf{f}^*|\mathbf{y}) \approx \mathcal{N}(\mathbf{f}^*|(\mathbf{K}_{M*}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{u}}, \mathbf{\Lambda}_* + \mathbf{K}_{M*}^T\mathbf{Q}_{MM}^{-1}\mathbf{K}_{M*})$$

- スパースガウス過程回帰モデルの対数周辺尤度の変分下界（ELBO）

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y}|\mathbf{0}_N, \mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} + \sigma^2\mathbf{I}) - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{\Lambda})$$

1．ガウス過程回帰モデル

通常のガウス過程回帰モデルには基本的に紹介だけに留めます。

前提は、観測誤差 $\epsilon \sim \mathcal{N}(0, \sigma^2)$ の観測データ $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$ が与えられていて、 y は平均が 0 になるように正規化してあるとすると、入力 $\mathbf{x}_n$ と出力 y_n の間に

$$y_n = f(\mathbf{x}_n) + \epsilon_n, \text{ where } f \sim \mathrm{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$$

の関係があるとします。ここで $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$ 、 $\mathbf{K}_{NN}$ を $K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'})$ を要素とする共分散行列とすると

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I})$$

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{NN})$$

と表せます。

ここで、新たな入力 $\mathbf{x}^*$ に対して、観測値の集合 y^* を考えます。新たな入力は1つでなくて良いので、それぞれ $\mathbf{X}^*$ 、 $\mathbf{y}^*$ とすれば、その予測分布は、

$$\begin{aligned} p(\mathbf{y}^*|\mathbf{X}^*, \mathcal{D}) = \mathcal{N}(\mathbf{y}^*|\mathbf{K}_{N*}^T(\sigma^2\mathbf{I} + \mathbf{K}_{NN})^{-1}\mathbf{y}, \\ \mathbf{K}_{**} - \mathbf{K}_{N*}^T(\sigma^2\mathbf{I} + \mathbf{K}_{NN})^{-1}\mathbf{K}_{N*}) \end{aligned}$$

で与えられます。

ハイパーパラメータの最適化については、 $\mathbf{y}$ の対数周辺尤度 $\log p(\mathbf{y}|\mathbf{X})$ をカーネル関数のパラメータと観測ノイズについて最大化します。

まず周辺尤度 $p(\mathbf{y}|\mathbf{X})$ は

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}, \mathbf{f}|\mathbf{X}) d\mathbf{f} \\ &= \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f} \\ &= \int \mathcal{N}(\mathbf{y} - \mathbf{f}|\mathbf{0}, \sigma^2\mathbf{I}) \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{NN}) d\mathbf{f} \end{aligned}$$

とかけて、これは正規分布同士の畳み込みであることがわかるので、

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \mathcal{N}(\mathbf{y}|\mathbf{0} + \mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{NN}) \\ &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{NN}) \end{aligned}$$

となります。

この対数をとって最適化に関係のある部分、つまりカーネル関数のハイパーパラメータと観測ノイズを含む項だけを残せばよいので

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{X}) &= \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{NN}) \\ &= \frac{1}{(2\pi)^{N/2}} \frac{1}{|\sigma^2\mathbf{I} + \mathbf{K}_{NN}|^{1/2}} \exp(-\frac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{I} + \mathbf{K}_{NN})^{-1}\mathbf{y}) \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2\mathbf{I} + \mathbf{K}_{NN}| - \frac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{I} + \mathbf{K}_{NN})^{-1}\mathbf{y} \\ &= -\log|\sigma^2\mathbf{I} + \mathbf{K}_{NN}| - \mathbf{y}^T(\sigma^2\mathbf{I} + \mathbf{K}_{NN})^{-1}\mathbf{y} + \text{ const} \\ &\propto -\log|\sigma^2\mathbf{I} + \mathbf{K}_{NN}| - \mathbf{y}^T(\sigma^2\mathbf{I} + \mathbf{K}_{NN})^{-1}\mathbf{y} \end{aligned}$$

となって最適化のための目的関数が得られました。

2. スパースガウス過程回帰モデル

補助変数法にはいくつかの流儀があるようですが、ここでのモデルは、Fully independent training conditional (FITC) approximation と呼ばれるものです。

2.0 前提

基本的に、１章での通常のガウス過程回帰の設定を引き継ぎます。ここで、 $f(\cdot)$ 定義域内に、 $M < N$ 個の補助入力点 $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)$ を配置し、 $\mathbf{Z}$ 上の $f(\cdot)$ による出力値 $\mathbf{u} = (f(\mathbf{z}_1), \ldots, f(\mathbf{z}_M))$ を導入し、これを補助変数ベクトルと呼びます。ここで、 $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{MM})$ となり、 $\mathbf{f}$ と $\mathbf{u}$ の同時分布 $p(\mathbf{f}, \mathbf{u})$ は、

$$p(\mathbf{f}, \mathbf{u}) = \mathcal{N}\left(\begin{pmatrix}\mathbf{f} \\ \mathbf{u}\end{pmatrix} \middle| \begin{pmatrix}\mathbf{0}_N \\ \mathbf{0}_M\end{pmatrix}, \begin{pmatrix}\mathbf{K}_{NN} & \mathbf{K}_{NM} \\ \mathbf{K}_{NM}^T & \mathbf{K}_{MM}\end{pmatrix}\right)$$

で与えられます。

補助変数法では、新しい入力点 $\mathbf{X}^*$ の予測分布に

$$\begin{aligned} p(\mathbf{f}^*|\mathbf{y}) &= \int p(\mathbf{f}^*|\mathbf{f}) p(\mathbf{f}|\mathbf{y}) d\mathbf{f} \\ &\approx \int p(\mathbf{f}^*|\mathbf{u}) p(\mathbf{u}|\mathbf{y}) d\mathbf{u} \end{aligned}$$

という近似を行い、この近似が精度良く成り立つなら、新しい入力点の代わりに観測データの入力点 $\mathbf{X}$ を入れても

$$p(\mathbf{f}|\mathbf{y}) \approx \int p(\mathbf{f}|\mathbf{u}) p(\mathbf{u}|\mathbf{y}) d\mathbf{u}$$

が成り立ちます。ここで 観測点 $\mathbf{y}$ の生成過程は

1. $\mathbf{u} \sim \mathcal{N}(\mathbf{0}_M, \mathbf{K}_{MM})$
2. $\mathbf{f}|\mathbf{u} \sim \mathcal{N}(\mathbf{K}_{MN}^T \mathbf{K}_{MM}^{-1} \mathbf{u}, \mathbf{\Lambda})$,
   $\text{where } \mathbf{\Lambda} = \text{ diag}(\mathbf{K}_{NN} - \mathbf{K}_{MN}^T \mathbf{K}_{MM}^{-1} \mathbf{K}_{MN})$
3. $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 \mathbf{I}_N)$

とします。

2.1 予測分布の導出

補助変数の事後確率 $p(\mathbf{u}|\mathbf{y})$

次に、補助変数の事後確率 $p(\mathbf{u}|\mathbf{y})$ を求めます。まず、ベイズの定理より

$$\begin{aligned} & p(\mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{u}) p(\mathbf{u})}{p(\mathbf{y})} \\ \iff & \ln p(\mathbf{u}|\mathbf{y}) = \ln \frac{p(\mathbf{y}|\mathbf{u}) p(\mathbf{u})}{p(\mathbf{y})} \\ \iff & \ln p(\mathbf{u}|\mathbf{y}) = \ln p(\mathbf{y}|\mathbf{u}) + \ln p(\mathbf{u}) - \ln p(\mathbf{y}) \end{aligned}$$

となります。 $p(\mathbf{y}|\mathbf{u})$ は畳み込みにより、

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{u}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})d\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{y}-\mathbf{f}|\mathbf{0},\sigma^2\mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f} \\
&= \mathcal{N}(\mathbf{y}|\mathbf{0}+\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\sigma^2\mathbf{I}+\mathbf{\Lambda}) \\
&= \mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\sigma^2\mathbf{I}+\mathbf{\Lambda})
\end{aligned}
$$

となります。 $p(\mathbf{y}|\mathbf{u})$ と $p(\mathbf{u})$ を $\ln p(\mathbf{u}|\mathbf{y})$ の左辺に代入して、 $\mathbf{u}$ に関して偏微分すると

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{u}|\mathbf{y})}{\partial \mathbf{u}} &= \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y}|\mathbf{u}) + \ln p(\mathbf{u}) - \ln p(\mathbf{y})\} \\
&= \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y}|\mathbf{u})\} + \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{u})\} - \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y})\} \\
&= \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y}|\mathbf{u})\} + \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{u})\}
\end{aligned}
$$

で、 $\frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y}|\mathbf{u})\}$ と $\frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{u})\}$ はそれぞれ、

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y}|\mathbf{u})\} &= \frac{\partial}{\partial \mathbf{u}}\{\ln \mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\sigma^2\mathbf{I}+\mathbf{\Lambda})\} \\
&= \frac{\partial}{\partial \mathbf{u}}\left\{-\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\sigma^2\mathbf{I}+\mathbf{\Lambda}| - \frac{1}{2}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})\right\} \\
&= -\frac{1}{2}\frac{\partial}{\partial \mathbf{u}}\left\{(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})\right\} \\
&= -\frac{1}{2}\frac{\partial}{\partial \mathbf{u}}\left\{\mathbf{y}^T(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} - \mathbf{y}^T(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u} - (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} + (\right. \\
&= -\frac{1}{2}\frac{\partial}{\partial \mathbf{u}}\left\{-2[\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y}\}^T\mathbf{u}]^T + \mathbf{u}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}\right\} \\
&= -\frac{1}{2}\left\{-2\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} + 2\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}\right\} \\
&= \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} - \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{u})\} &= \frac{\partial}{\partial \mathbf{u}}\{\ln \mathcal{N}(\mathbf{u}|\mathbf{0}_M,\mathbf{K}_{MM})\} \\
&= \frac{\partial}{\partial \mathbf{u}}\left\{-\frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{K}_{MM}| - \frac{1}{2}\mathbf{u}^T\mathbf{K}_{MM}^{-1}\mathbf{u}\right\} \\
&= -\frac{1}{2}\frac{\partial}{\partial \mathbf{u}}\left(\mathbf{u}^T\mathbf{K}_{MM}^{-1}\mathbf{u}\right) \\
&= -\mathbf{K}_{MM}^{-1}\mathbf{u}
\end{aligned}
$$

となので、

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{u}|\mathbf{y})}{\partial \mathbf{u}} &= \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{y}|\mathbf{u})\} + \frac{\partial}{\partial \mathbf{u}}\{\ln p(\mathbf{u})\} \\
&= \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} - \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u} - \mathbf{K}_{MM}^{-1}\mathbf{u} \\
&= -\{\mathbf{K}_{MM}^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}^{-1} + \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\}\mathbf{u} + \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -[\mathbf{K}_{MM}^{-1}\{\mathbf{K}_{MM} + \mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T\}\mathbf{K}_{MM}^{-1}]\mathbf{u} + \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -(\mathbf{K}_{MM}^{-1}\mathbf{Q}_{MM}\mathbf{K}_{MM}^{-1})\mathbf{u} + \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -\widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{u} + \mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -\widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{u} + \widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{u}}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -\widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{u} + \widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -\widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{u} + \widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y} \\
&= -\widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\mathbf{u} + \widehat{\mathbf{\Sigma}}_{\mathbf{u}}^{-1}\widehat{\mathbf{u}},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{Q}_{MM} &= \mathbf{K}_{MM} + \mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T \\
\widehat{\mathbf{\Sigma}}_{\mathbf{u}} &= \mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MM} \\
\widehat{\mathbf{u}} &= \mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{y}
\end{aligned}
$$

となります。一般の多変量ガウス分布の場合と見比べれば、

$$
p(\mathbf{u}|\mathbf{y}) = \mathcal{N}(\widehat{\mathbf{u}}, \widehat{\mathbf{\Sigma}}_{\mathbf{u}})
$$

となって補助変数の事後確率が得られました。

予測分布 $p(\mathbf{f}^*|\mathbf{y})$

ここで $p(\mathbf{f}|\mathbf{y})$ と $p(\mathbf{f}^*|\mathbf{y})$ の近似分布に戻ります。

まず $p(\mathbf{f}|\mathbf{y}) \approx \int p(\mathbf{f}|\mathbf{u})p(\mathbf{u}|\mathbf{y})d\mathbf{u}$ について、

$$
\mathbf{u}|\mathbf{y} \sim \mathcal{N}(\mathbf{u}|\widehat{\mathbf{u}}, \widehat{\mathbf{\Sigma}}_{\mathbf{u}})
$$

$$
\mathbf{f}|\mathbf{u} \sim \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{u} + \mathbf{0}_N, \mathbf{\Lambda})
$$

なので、多変量ガウス分布の線形変換より

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{y}) &\approx \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{u}} + \mathbf{0}_N, \mathbf{\Lambda} + (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{\Sigma}}_{\mathbf{u}}(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})^T) \\
&= \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{u}}, \mathbf{\Lambda} + (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{\Sigma}}_{\mathbf{u}}(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})^T) \\
&= \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{u}}, \mathbf{\Lambda} + (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MM}^T(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})^T) \\
&= \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{u}}, \mathbf{\Lambda} + \mathbf{K}_{MN}^T\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MN})
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{\Lambda} &= \ \mathrm{diag}(\mathbf{K}_{NN} - \mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}) \\
\mathbf{Q}_{MM} &= \mathbf{K}_{MM} + \mathbf{K}_{MN}(\sigma^2\mathbf{I} + \mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T \\
\widehat{\mathbf{u}} &= \mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I} + \mathbf{\Lambda})^{-1}\mathbf{y}
\end{aligned}
$$

となります。

$p(\mathbf{f}^*|\mathbf{y}) \approx \int p(\mathbf{f}^*|\mathbf{u})p(\mathbf{u}|\mathbf{y})d\mathbf{u}$ も同様に考えれば、

$$
p(\mathbf{f}^*|\mathbf{y}) \approx \mathcal{N}(\mathbf{f}^*|(\mathbf{K}_{M*}^T\mathbf{K}_{MM}^{-1})\widehat{\mathbf{u}}, \mathbf{\Lambda}_* + \mathbf{K}_{M*}^T\mathbf{Q}_{MM}^{-1}\mathbf{K}_{M*})
$$

where

$$
\begin{aligned}
\mathbf{\Lambda}_* &= \ \mathrm{diag}(\mathbf{K}_{**} - \mathbf{K}_{M*}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{M*}) \\
\mathbf{Q}_{MM} &= \mathbf{K}_{MM} + \mathbf{K}_{MN}(\sigma^2\mathbf{I} + \mathbf{\Lambda})^{-1}\mathbf{K}_{MN}^T \\
\widehat{\mathbf{u}} &= \mathbf{K}_{MM}\mathbf{Q}_{MM}^{-1}\mathbf{K}_{MN}(\sigma^2\mathbf{I} + \mathbf{\Lambda})^{-1}\mathbf{y}
\end{aligned}
$$

と予測分布が求まります。

2.2 ハイパーパラメータの最適化

周辺尤度 $p(\mathbf{y})$

ハイパーパラメータの最適化にそのまま使用するわけではありませんが、周辺尤度 $p(\mathbf{y})$ 、

$$
p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}
$$

つまり補助変数法の確率的生成モデルのエビデンスについても求めておきます。まずは

$$
\begin{aligned}
\mathbf{u} &\sim \mathcal{N}(\mathbf{u}|\mathbf{0}_M, \mathbf{K}_{MM}) \\
\mathbf{f}|\mathbf{u} &\sim \mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}, \mathbf{\Lambda}) = \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{u} + \mathbf{0}_N, \mathbf{\Lambda})
\end{aligned}
$$

を見ると、 $\mathbf{f}|\mathbf{u}$ はアフィン変換 $(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{u} + \mathbf{0}_N$ を平均とした精度 $\mathbf{\Lambda}$ の多変量ガウス分布に従っているので、多変量ガウス分布の線形変換により

$$
\begin{aligned}
p(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \\
&= \mathcal{N}(\mathbf{f}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{0}_M + \mathbf{0}_N, \mathbf{\Lambda} + (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{K}_{MM}(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})^T) \\
&= \mathcal{N}(\mathbf{f}|(\mathbf{0}_N, \mathbf{\Lambda} + \mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN})
\end{aligned}
$$

が得られ、畳み込みにより

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})\mathcal{N}(\mathbf{f}|(\mathbf{0}_N,\mathbf{\Lambda}+\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN})d\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{y}-\mathbf{f}|\mathbf{0}_N,\sigma^2\mathbf{I})\mathcal{N}(\mathbf{f}|(\mathbf{0}_N,\mathbf{\Lambda}+\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN})d\mathbf{f} \\
&= \mathcal{N}(\mathbf{y}|\mathbf{0}_N+\mathbf{0}_N,\sigma^2\mathbf{I}+\mathbf{\Lambda}+\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}) \\
&= \mathcal{N}(\mathbf{y}|\mathbf{0}_N,\sigma^2\mathbf{I}+\mathbf{\Lambda}+\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN})
\end{aligned}
$$

となり、目的のエビデンスが得られます

ELBO

最適化の場合にはこちらを目的関数として用います。先ほど上で考えた周辺尤度の対数をとれば

$$
\log p(\mathbf{y}|\mathbf{X}) = \log \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}
$$

となり、ここから導出したいのはこれを下から抑える表現です。まず、イェンセンの不等式を用いて

$$
\begin{aligned}
\log p(\mathbf{y}|\mathbf{u}) &= \log \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})d\mathbf{f} \\
&\geq \int \log\{p(\mathbf{y}|\mathbf{f})\}p(\mathbf{f}|\mathbf{u})d\mathbf{f} =: \mathcal{L}_1 \\
&= \int \left\{-\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| - \frac{1}{2}(\mathbf{y}-\mathbf{f})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{f})\right\}\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f} \\
&= \int \left(-\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}|\right)\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f} + \int \left\{-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{f})\right\}\mathcal{N}(\mathbf{f}|\mathbf{K} \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| - \frac{1}{2}\int (\mathbf{f}-\mathbf{y})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{f}-\mathbf{y})\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f}
\end{aligned}
$$

ここで、 $\int (\mathbf{f}-\mathbf{y})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{f}-\mathbf{y})\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f}$ は二次形式の期待値で、

$$
\begin{aligned}
&\int (\mathbf{f}-\mathbf{y})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{f}-\mathbf{y})\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f} \\
&\qquad = \mathrm{tr}((\sigma^2\mathbf{I})^{-1}\mathbf{\Lambda}) + (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}-\mathbf{y})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}-\mathbf{y}) \\
&\qquad = \frac{1}{\sigma^2}\mathrm{tr}(\mathbf{\Lambda}) + (\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})
\end{aligned}
$$

となるので、

$$
\begin{aligned}
\mathcal{L}_1 &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| - \frac{1}{2}\int (\mathbf{f}-\mathbf{y})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{f}-\mathbf{y})\mathcal{N}(\mathbf{f}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\mathbf{\Lambda})d\mathbf{f} \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| - \frac{1}{2}\left\{\frac{1}{\sigma^2}\mathrm{tr}(\mathbf{\Lambda}) + (\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})\right\} \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda}) - \frac{1}{2}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}) \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| - \frac{1}{2}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda}) \\
&= \log\mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u},\sigma^2\mathbf{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda})
\end{aligned}
$$

となります。最初の対数周辺尤度に戻って、得られた $p(\mathbf{y}|\mathbf{u})$ の下界を代入すると対数周辺尤度の下界の表現が得られて、これを ELBO とします。

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &= \log \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u} \\ &= \log \int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \\ &\geq \log \int \exp(\mathcal{L}_1)p(\mathbf{u})d\mathbf{u} =: \mathcal{L}_2\end{aligned}$$

$$\begin{aligned}\mathcal{L}_2 &= \log \int \exp(\mathcal{L}_1)p(\mathbf{u})d\mathbf{u} \\ &= \log \int \mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}, \sigma^2\mathbf{I}) \exp\left\{-\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda})\right\}\mathcal{N}(\mathbf{u}|\mathbf{0}_M, \mathbf{K}_{MM})d\mathbf{u} \\ &= \log \int \mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{u}|\mathbf{0}_M, \mathbf{K}_{MM})d\mathbf{u} - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda})\end{aligned}$$

ここで、周辺尤度を求めたときに使用したアフィン変換のロジックをもう一度使えば

$$\begin{aligned}&\int \mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{u}|\mathbf{0}_M, \mathbf{K}_{MM})d\mathbf{u} \\ &\quad= \int \mathcal{N}(\mathbf{y}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{u} + \mathbf{0}_N, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{u}|\mathbf{0}_M, \mathbf{K}_{MM})d\mathbf{u} \\ &\quad= \mathcal{N}(\mathbf{y}|(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{0}_M + \mathbf{0}_N, \sigma^2\mathbf{I} + (\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})\mathbf{K}_{MM}(\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1})^T) \\ &\quad= \mathcal{N}(\mathbf{y}|\mathbf{0}_N, \mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} + \sigma^2\mathbf{I})\end{aligned}$$

となるので、これを代入して、

$$\begin{aligned}\mathcal{L}_2 &= \log \int \mathcal{N}(\mathbf{y}|\mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{u}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{u}|\mathbf{0}_M, \mathbf{K}_{MM})d\mathbf{u} - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda}) \\ &= \log \mathcal{N}(\mathbf{y}|\mathbf{0}_N, \mathbf{K}_{MN}^T\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} + \sigma^2\mathbf{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{\Lambda})\end{aligned}$$

となり、これで ELBO の導出は完了です。

3. 数学の補足

3.1 多変量正規分布の対数とその微分

D 次元の多変量正規分布の確率密度関数は、

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

となので、対数をとると

$$\begin{aligned}\ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) &= \ln\left[\frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right] \\ &= -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\end{aligned}$$

となる。これを $\mathbf{x}$ について微分すると

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{x}}\{\ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\} &= \frac{\partial}{\partial \mathbf{x}}\left\{-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
&= \frac{\partial}{\partial \mathbf{x}}\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
&= \frac{\partial}{\partial \mathbf{x}}\left\{-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \\
&= -\frac{1}{2}\left\{(\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^T)\mathbf{x} - 2(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1})^T\right\} \\
&= -\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}
\end{aligned}
$$

となる。確率変数が多変量正規分布に従うことが分かっているとき、確率密度関数全体を計算する代わりに部分的に計算して一般の式と比較すると楽に望む結果が得られることがある。

3.2 多変量ガウス分布の畳み込み

確率分布の畳み込み

確率分布が連続の場合のみを紹介します。まず、２つの独立な確率分布 X と Y から、 $Z = X + Y$ という確率分布を考えます。ここで Z の累積分布関数は、

$$
\begin{aligned}
F_Z(z) &= P(Z \leq z) \\
&= P(X + Y \leq z) \\
&= \int_{x \in \Omega_X} P(X + Y \leq z | X = x) f_X(x) dx \\
&= \int_{x \in \Omega_X} P(Y \leq z - x | X = x) f_X(x) dx \\
&= \int_{x \in \Omega_X} P(Y \leq z - x) f_X(x) dx \quad (\because X \perp\!\!\!\perp Y) \\
&= \int_{x \in \Omega_X} F_Y(z - x) f_X(x) dx
\end{aligned}
$$

となるので、 z について微分すれば

$$
\begin{aligned}
f_Z(z) &= \frac{d}{dz} F_Z(z) \\
&= \int_{x \in \Omega_X} f_X(x) f_Y(z - x) dx
\end{aligned}
$$

となり、新たな確率変数 Z の確率密度関数が得られました。

多変量ガウス分布の畳み込み

２つの独立な D 次元の多変量正規分布 $\mathbf{X}$ と $\mathbf{Y}$ を考えます。その確率密度関数はそれぞれ

$$
\begin{aligned}
\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_1|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} \\
\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_2|}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{y}-\boldsymbol{\mu}_2)\right\}
\end{aligned}
$$

として、新たな確率変数 $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ の従う確率密度関数を導出します。

まずは、一般の式から変形していきます。

$$
\begin{aligned}
f_{\mathbf{Z}}(\mathbf{z}) &= \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{x} \\
&= \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{z}-\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}(\mathbf{z}-\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) d\mathbf{x} \\
&= \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_1|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} \cdot \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_2|}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{z}-\right. \\
&= \frac{1}{\sqrt{(2\pi)^D}} \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + (\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2)\right)\right.
\end{aligned}
$$

ここで、 $\mathbf{A}$ と $\mathbf{B}$ が対称行列のとき、

$$
\begin{aligned}
&(\mathbf{x}-\mathbf{a})^T\mathbf{A}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})^T\mathbf{B}(\mathbf{x}-\mathbf{b}) \\
&\qquad = (\mathbf{x}-\mathbf{c})^T(\mathbf{A}+\mathbf{B})(\mathbf{x}-\mathbf{c}) + (\mathbf{a}-\mathbf{b})^T\mathbf{C}(\mathbf{a}-\mathbf{b})
\end{aligned}
$$

where

$$
\mathbf{c} = (\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})
$$

$$
\mathbf{C} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} = (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}
$$

なので、

$$
\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + (\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2) \\
&\qquad = \left\{\mathbf{x} - (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right\}^T \left\{(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}\right\}^{-1} \left\{\mathbf{x} - (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}\right. \\
&\qquad\quad + \{\mathbf{z} - (\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\}^T (\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} \{\mathbf{z} - (\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\}
\end{aligned}
$$

となり、

$$
\begin{aligned}
f_{\mathbf{Z}}(\mathbf{z}) &= \frac{1}{\sqrt{(2\pi)^D}} \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_1||\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+(\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{z}-\mathbf{x}-\boldsymbol{\mu}_2)\right.\right. \\
&= \frac{\sqrt{|\mathbf{\Sigma}_1+\mathbf{\Sigma}_2|}}{\sqrt{(2\pi)^D}\sqrt{|\mathbf{\Sigma}_1+\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}^T\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)^{-1}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}\right\} \\
&\quad \times \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_1||\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{x}-(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right\}^T\left\{(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}\right\}^{-1}\right. \\
&= \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_1+\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}^T\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)^{-1}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}\right\} \\
&\quad \times \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \frac{1}{\sqrt{(2\pi)^D}} \sqrt{\frac{|\mathbf{\Sigma}_1+\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1||\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{x}-(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right\}^T\left\{(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-}\right.\right. \\
&= \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_1+\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}^T\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)^{-1}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}\right\} \\
&\quad \times \int_{\mathbf{x}\in\Omega_{\mathbf{X}}} \frac{1}{\sqrt{(2\pi)^D|(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{x}-(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\right\}^T\left\{(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-}\right.\right. \\
&= \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_1+\mathbf{\Sigma}_2|}} \exp\left\{-\frac{1}{2}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}^T\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)^{-1}\left\{\mathbf{z}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right\}\right\} \\
&= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_1+\boldsymbol{\mu}_2, \mathbf{\Sigma}_1+\mathbf{\Sigma}_2)
\end{aligned}
$$

となって、 $\mathbf{Z}$ の確率密度関数が得られた。

3.3 多変量ガウス分布の線形変換

平均 $\boldsymbol{\mu}$ 、共分散 $\mathbf{\Sigma}_x$ をもつガウス分布に従う確率変数 $\mathbf{x}$ に対して、アフィン変換 $\mathbf{W}\mathbf{x}+\mathbf{b}$ を平均とした精度 $\mathbf{\Sigma}_y$ のガウス分布に従う確率変数 $\mathbf{y}$ を考えます。すなわち、

$$
p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}_x)
$$

$$
p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{x}+\mathbf{b}, \mathbf{\Sigma}_y)
$$

とするとき、 $p(\mathbf{y})$ は

$$
p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{W}\boldsymbol{\mu}+\mathbf{b}, \mathbf{\Sigma}_y+\mathbf{W}\mathbf{\Sigma}_x\mathbf{W}^T)
$$

となります。これの証明については、参照した「ベイズ深層学習」に載っている以上のことが書けなかったので与えないことにします。他には下のサイトにも載っています。

Linear Transformation of Gaussian Random Variable

3.4 共分散行列が零行列の多変量ガウス分布

たまに、計算をしているとガウス分布の分散が 0 になってしまう時があります。例えば、ガウス過程回帰モデルの観測ノイズをカーネル関数に含めた形で表すと、

$$
y_n = f(\mathbf{x}_n),\ \text{where}\ f \sim \mathrm{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}') + \sigma^2\delta(n, n'))
$$

で、この周辺尤度を考えると、

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}, \mathbf{f}|\mathbf{X}) d\mathbf{f} \\
&= \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{0}) \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{NN} + \sigma^2 \mathbf{I}) d\mathbf{f}
\end{aligned}
$$

となって、 $\mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{0})$ のように共分散行列が $\mathbf{0}$ のガウス分布が出てきます。ガウス分布の確率密度関数を考えると、共分散行列の逆行列を求める操作が含まれるのでこれはおかしいのですが、 $\mathbf{y} = \mathbf{f}$ のときに必ず 1 を返すと形式的にみなすことで、ある種の正規分布として考えられます。すると

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) &= \int \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{0}) \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{NN} + \sigma^2 \mathbf{I}) d\mathbf{f} \\
&= \mathcal{N}(\mathbf{y}|\mathbf{0} + \mathbf{0}, \mathbf{0} + \mathbf{K}_{NN} + \sigma^2 \mathbf{I}) \\
&= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{NN} + \sigma^2 \mathbf{I})
\end{aligned}
$$

となって、ノイズを含めない形の $\mathbf{y}$ の周辺尤度と同じものが得られました。

このような確率変数を “degenerate” あるいは、 “deterministic” などと言ったりするようです。

3.5 二次形式の期待値

期待値 $\boldsymbol{\mu}$ 、共分散行列 $\boldsymbol{\Sigma}$ の確率変数 $\mathbf{X}$ （n x 1 の確率変数ベクトル）があるとして、二次形式 $(\mathbf{X} - \mathbf{a})^T \mathbf{A} (\mathbf{X} - \mathbf{a})$ の期待値を考えます。ここで、 $\mathbf{A}$ は対象行列なので、

$$
(\mathbf{X} - \mathbf{a})^T \mathbf{A} (\mathbf{X} - \mathbf{a}) = \mathbf{X}^T \mathbf{A} \mathbf{X} - 2\mathbf{a}^T \mathbf{A} \mathbf{X} + \mathbf{a}^T \mathbf{A} \mathbf{a}
$$

となり、

$$
\begin{aligned}
\mathbb{E}[(\mathbf{X} - \mathbf{a})^T \mathbf{A} (\mathbf{X} - \mathbf{a})] &= \mathbb{E}[\mathbf{X}^T \mathbf{A} \mathbf{X} - 2\mathbf{a}^T \mathbf{A} \mathbf{X} + \mathbf{a}^T \mathbf{A} \mathbf{a}] \\
&= \mathbb{E}[\mathbf{X}^T \mathbf{A} \mathbf{X}] - 2\mathbf{a}^T \mathbf{A} \mathbb{E}[\mathbf{X}] + \mathbf{a}^T \mathbf{A} \mathbf{a} \mathbb{E}[1] \quad (\because \text{期待値の線形性}) \\
&= \mathbb{E}[\mathbf{X}^T \mathbf{A} \mathbf{X}] - 2\mathbf{a}^T \mathbf{A} \boldsymbol{\mu} + \mathbf{a}^T \mathbf{A} \mathbf{a}
\end{aligned}
$$

で、 $\mathbb{E}[\mathbf{X}^T \mathbf{A} \mathbf{X}]$ は、

$$
\begin{aligned}
\mathbb{E}[\mathbf{X}^T \mathbf{A} \mathbf{X}] &= \mathbb{E}[\mathrm{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X})] \quad (\because \mathbf{X}^T \mathbf{A} \mathbf{X} \text{はスカラー}) \\
&= \mathbb{E}[\mathrm{tr}(\mathbf{A} \mathbf{X} \mathbf{X}^T)] \quad (\because \mathrm{tr}(ABC) = \mathrm{tr}(BCA)) \\
&= \mathrm{tr}(\mathbf{A} \mathbb{E}[\mathbf{X} \mathbf{X}^T]) \quad (\because \mathbb{E}[\mathrm{tr}(A)] = \mathrm{tr}(\mathbb{E}[A])) \\
&= \mathrm{tr}(\mathbf{A} [\mathrm{Cov}(\mathbf{X}, \mathbf{X}) + \mathbb{E}[\mathbf{X}] \mathbb{E}[\mathbf{X}]^T) \\
&= \mathrm{tr}(\mathbf{A} (\boldsymbol{\Sigma} + \boldsymbol{\mu} \boldsymbol{\mu}^T)) \\
&= \mathrm{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \mathrm{tr}(\mathbf{A} \boldsymbol{\mu} \boldsymbol{\mu}^T) \\
&= \mathrm{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}) \\
&= \mathrm{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}
\end{aligned}
$$

なので、最終的には

$$\begin{aligned}\mathbb{E}[(\mathbf{X}-\mathbf{a})^T\mathbf{A}(\mathbf{X}-\mathbf{a})] &= \mathbb{E}[\mathbf{X}^T\mathbf{A}\mathbf{X}] - 2\mathbf{a}^T\mathbf{A}\boldsymbol{\mu} + \mathbf{a}^T\mathbf{A}\mathbf{a} \\ &= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - 2\mathbf{a}^T\mathbf{A}\boldsymbol{\mu} + \mathbf{a}^T\mathbf{A}\mathbf{a} \\ &= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\mathbf{a})^T\mathbf{A}(\boldsymbol{\mu}-\mathbf{a})\end{aligned}$$

となります。

4. まとめ

この記事では、ガウス過程回帰の計算コスト削減アルゴリズムである補助変数法について紹介しました。当然のことながら、この記事で紹介したことは補助変数法のすべてではありません。ELBO をグローバル変数を明示的な形で定義することによって、データをミニバッチに分けてのハイパーパラメータの最適化が可能になり、より巨大なデータに立ち向かうことができるようになるようです。他にも「ガウス過程と機械学習（講談社）」では KISS-GP なる手法も紹介されています。

ガウス過程回帰に限らず、機械学習のアルゴリズム全般は、広い範囲の数学を用いて記述されているため、初学者にとってはなかなか勉強するのが大変だと思います。この記事が少しでもそんな人の助けになれば光栄です。

参照

1. **ガウス過程と機械学習**
2. **ベイズ深層学習**
3. **ベクトル・行列を含む微分**
4. **Sparse Gaussian process**
5. **Derivation of SGPR equation**
6. **Sparse Gaussian Processes using Pseudo-inputs**
7. **Variational Learning of Inducing Variables in Sparse Gaussian Process**
8. **Variational Model Selection for Sparse Gaussian Process Regression**
9. **A Unifying View of Sparse Approximate Gaussian Process Regression**
10. **Gaussian Processes for Big Data**
11. **Convolution integrals of Normal distribution functions**
12. **Expectation of a quadratic form**
13. **Degenerate distribution**
14. **Linear Transformation of Gaussian Random Variable**