贝叶斯稀疏多项式混沌展开(BSPCE)的思路与推导

Author: Jian Liu

Time: 2020-02-26

Abstract:

1. 第一章介绍了Sobol' 分解、方差分析。
2. 第二章介绍了多项式混沌展开，并给出了其重要应用：方差分解。
3. 第三章为 BSPCE 公式的详细推导。主要包括两个方面：(1) 参数估计、(2) 模型选择。参数估计就是给定基函数后对基函数系数进行估计，方法为最大后验估计，目标是使模型更好地拟合样本数据。模型选择则是对模型所包含的基函数进行评估，评估指标为 KIC，目标是选择泛化能力强的模型。
4. 第四章为 BSPCE 的构建算法，核心点为两次排序及基函数的扩充策略。此外，还对 BSPCE 的进一步研究进行了展望。
5. 第五章给出了关于模型选择指标的一些参考文献及相关结论。
6. 附录给出了使用概率模型进行回归的方法。
7. 参考文献

贝叶斯稀疏多项式混沌展开(BSPCE)的思路与推导
1 Sobol' decomposition 与 ANOVA
2 Polynomial chaos expasion
3 模型选择与参数估计
3.1 参数估计
3.1.1 $\boldsymbol{a}$ 的后验估计
3.1.2 σ_ε^2 的后验估计
3.1.3 $\hat{\boldsymbol{a}}$ 和 $\hat{\sigma}_\varepsilon^2$ 的迭代计算
3.1.4 似然函数的计算
3.2 模型选择
3.2.1 KIC (Kashyap information criterion) 的推导
3.2.2 BSPCE 的 KIC
3.3 模型选择与参数估计的辩证关系
4 BSPCE 的构建算法及展望
4.1 BSPCE 的构建算法
4.2 BSPCE 算法的展望
5 文献阅读之 KIC、BIC、AIC
附录 - 概率回归
参考文献

1 Sobol' decomposition 与 ANOVA

Sobol' 分解：若函数 $M(\boldsymbol{x})$ ($x_i \in [0,1],\ i=1,\cdots,n$)**可积**，则 $M(\boldsymbol{x})$ 可分解为维数不断增加的各项之和:

$$M(\boldsymbol{x}) = M_0 + \sum_{i_1=1}^{n} M_{i_1}(x_{i_1}) + \sum_{i_2>i_1}^{n} M_{i_1 i_2}(x_{i_1}, x_{i_2}) + \cdots + M_{12\ldots n}(\boldsymbol{x})$$

其中，每个被加项对其任意一个自变量的积分为零，即:

$$\int_0^1 M_{i_1\ldots i_t}(x_{i_1}, \ldots, x_{i_t}) dx_{i_k} = 0, \quad 1 \leqslant k \leqslant t$$

这样的分解具有**唯一性**，各被加项的解析计算公式如下:

$$M_0 = \int_{\mathbb{K}^n} M(\boldsymbol{x}) d\boldsymbol{x}$$

$$M_{i_1}(x_{i_1}) = \int_{\mathbb{K}^{n-1}} M(\boldsymbol{x}) d\boldsymbol{x}_{\sim i_1} - M_0$$

$$M_{i_1 i_2}(x_{i_1}, x_{i_2}) = \int_{\mathbb{K}^{n-2}} M(\boldsymbol{x}) d\boldsymbol{x}_{\sim i_1, i_2} - M_{i_1}(x_{i_1}) - M_{i_2}(x_{i_2}) - M_0$$

$$\cdots$$

进一步地，若 $M(\boldsymbol{x})$ **平方可积**，则可推得各被加项存在如下的正交性：

$$\int_{\mathbb{K}^n} M_{i_1 \ldots i_k}(x_{i_1}, \ldots, x_{i_k}) M_{j_1 \ldots j_t}(x_{j_1}, \ldots, x_{j_t}) d\boldsymbol{x} = 0 \quad \text{for} \quad \{i_1 \ldots i_k\} \neq \{j_1 \ldots j_t\}$$

接下来，将输入参数看做随机变量，并假设 $x_i \sim \mathcal{U}(0,1)$ **（**$i = 1, \cdots, n$**），且相互独立**，则输出 $y = M(\boldsymbol{x})$ 也是随机变量，并可得方差分解公式：

$$\begin{aligned} D &= \int_{\mathbb{K}^n} M^2(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} - \left(\int_{\mathbb{K}^n} M(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \right)^2 \\ &= \int_{\mathbb{K}^n} M^2(\boldsymbol{x}) d\boldsymbol{x} - M_0^2 \\ &= \sum_{i_1=1}^{n} D_{i_1} + \sum_{i_2 > i_1}^{n} D_{i_1 i_2} + \cdots + D_{12 \ldots n} \end{aligned}$$

其中，偏方差 $D_{i_1 \ldots i_t}$：

$$D_{i_1 i_2 \ldots i_t} = \int_{\mathbb{K}^s} M^2_{i_1 \ldots i_t}(x_{i_1}, \ldots, x_{i_t}) dx_{i_1} \ldots dx_{i_t}$$

Sobol'指数：

$$S_{i_1 \ldots i_t} = \frac{D_{i_1 \ldots i_t}}{D} \in [0,1]$$

Sobol' 法就是一种基于 Sobol' 分解的蒙特卡洛法，通过采样计算有关 $M(\boldsymbol{x})$ 的积分，进而计算方差，得到 Sobol' 指数。此外，**若** x_i **不服从均匀分布，我们可以通过概率分布转换，将任意概率密度转化为** $[0,1]$ **上的均匀分布**：令 $z_i = F_i(x_i)$（$i = 1, 2, \cdots, n$），其中 $F_i(\cdot)$ 表示随机变量 x_i 的累积分布函数 (CDF)。由概率论中的概率密度变换可知，随机变量 $\boldsymbol{z} = (z_1, z_2, \cdots, z_n)$ 服从 $\mathbb{K}^n$ 上的均匀分布。**最后，我们需要指出的是，Sobol' 分解是对原函数最彻底的分解，除 Sobol' 分解外，更一般的还有 Hoeffding–Sobol' decomposition。**

2 Polynomial chaos expasion

多项式混沌理论起源于1938年 Wiener N．的文章《The homogeneous chaos》，并于1991年在 Ghanem RG, Spanos PD 的书《Stochastic finite elements—a spectral approach》中再次被应用到工程领域。需要说明的是，这里的“**混沌**”和动力系统中的“**混沌**”在概念上是有所区别的。前者表示一种随机性，后者表示由系统的非线性所导致的响应对初始条件的极端敏感性，比如蝴蝶效应。

随机变量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 各分量独立同分布，概率密度函数记为 $p(\boldsymbol{x})$，响应 $y = M(\boldsymbol{x})$ 可被多项式基函数以如下的形式展开：

$$y = M(\boldsymbol{x}) = \sum_{\boldsymbol{b} \in \mathbb{N}^n} a_{\boldsymbol{b}} \Psi_{\boldsymbol{b}}(\boldsymbol{x}), \quad \Psi_{\boldsymbol{b}}(\boldsymbol{x}) = \Psi_{b_1 \ldots b_n}(\boldsymbol{x}) = \prod_{i=1}^{n} \psi_{b_i}(x_i)$$

其中，$\boldsymbol{b} = b_1 \cdots b_n$（$b_i \in \mathbb{N}, 1 \leqslant i \leqslant n$）是一个 n 维索引，$a_{\boldsymbol{b}}$ 为多项式基函数 $\Psi_{\boldsymbol{b}}(\boldsymbol{x})$ 对应的系数，$\Psi_{\boldsymbol{b}}(\boldsymbol{x})$ 为 n 个单变量多项式函数 $\psi_{b_i}(x_i)$（$i = 1, \cdots, n$）的累积，下标的索引分量 b_i 给出了 $\psi_{b_i}(x_i)$ 的阶数，$\psi_{b_i}(x_i)$ 的具体形式与概率密度函数 $p(\boldsymbol{x})$ 的类型相关，并使多项式基函数满足如下的**加权**正交性：

$$\int \Psi_{\boldsymbol{b}}(\boldsymbol{x}) \Psi_{\boldsymbol{b}'}(\boldsymbol{x}) p(\boldsymbol{x}) = 0, \quad \text{for} \boldsymbol{b} \neq \boldsymbol{b}'$$

不同概率分布对应多项式 $\psi_{b_i}(x_i)$ 的类型如下：

1. 均匀分布对应 Legendre 多项式；
2. 高斯分布对应 Hermite 多项式;
3. 伽马分布对应 Laguerre 多项式；
4. 贝塔分布对应 Jacobi 多项式；

等等。对于其他情况，可以采用等概率转换等方法，先将原概率分布转换或近似为上述已知的情况，再进行展开。有很多工作都在研究这一问题，其中，文献 [4,5] 讨论了任意概率分布、输入分量间不独立、广义多项式混沌等内容。若已知多项式混沌展开，可在此基础上构建模型响应的 PDF 和 CDF，比如文献 [6]。文献 [7] 中采用最小角回归法 (least angle regression, LAR) 构建稀疏多项式混沌展开，并使用交叉验证的留一法 (LOO) 进行模型选取。

我们这里假设 $x_i \sim \mathcal{U}[0,1]$（$i=1,\cdots,n$），相应地各阶 Legendre 多项式：

$$\psi_0(x)=1,\ \psi_1(x)=\sqrt{3}(2x-1),\ \psi_2(x)=\frac{3\sqrt{5}}{2}(2x-1)^2-\frac{\sqrt{5}}{2},\cdots$$

可以看到，多项式混沌展开的定义中存在无穷项，而在实际计算中，多项式混沌展开通常被截断而只保留有限项。一个总阶数 $|\boldsymbol{b}|\equiv\sum_{i=1}^n b_i$ 不超过给定阶数 p 的被截断的多项式混沌展开如下：

$$y\simeq M_p(\boldsymbol{x})\equiv\sum_{\boldsymbol{b}\in\mathcal{A}^p}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(\boldsymbol{x}),\ \ \mathcal{A}^p\equiv\{\boldsymbol{b}\in\mathbb{N}^n:|\boldsymbol{b}|\leqslant p\}$$

构建多项式混沌展开所面临的主要问题是：1) 保留哪些基函数项，也就是模型选择的问题; 2) 基函数系数的求解，也就是参数估计的问题。记 $\mathcal{A}\subset\mathcal{A}^p$ 为多维索引 $\boldsymbol{b}$ 的一个有限子集，可知，给定 $\mathcal{A}$ 就给定了保留哪些基函数项。若此时基函数系数也已知，则多项式混沌展开 $M_{\mathcal{A}}(\boldsymbol{x})=\sum_{\boldsymbol{b}\in\mathcal{A}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(\boldsymbol{x})$ 也就已知了。**现基于多项式混沌展开来推导方差分解。**首先定义记号 $I_{i_1\ldots i_t}$：

$$I_{i_1\ldots i_t}=\left\{\boldsymbol{b}\in\mathcal{A}:\ \ \forall k=1,\ldots n\ \begin{array}{ll}k\in(i_1\ldots i_t), & b_k>0\\ k\notin(i_1\ldots i_t), & b_k=0\end{array}\right\}$$

借助符号 $I_{i_1\ldots i_t}$ 改写 $M_{\mathcal{A}}(\boldsymbol{x})$：

$$\begin{aligned}M_{\mathcal{A}}=a_{\boldsymbol{0}}&+\sum_{i_1=1}^n\sum_{\boldsymbol{b}\in I_{i_1}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(x_{i_1})+\sum_{i_2>i_1}^n\sum_{\boldsymbol{b}\in I_{i_1i_2}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(x_{i_1},x_{i_2})\\&+\cdots+\sum_{i_t>\cdots>i_1}^n\sum_{\boldsymbol{b}\in I_{i_1\ldots i_t}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(x_{i_1},\ldots,x_{i_t})\\&+\cdots+\sum_{\boldsymbol{b}\in I_{1,\ldots,n}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(\boldsymbol{x})\end{aligned}$$

可以看到这里的 $\sum_{\boldsymbol{b}\in I_{i_1\ldots i_t}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(x_{i_1},\ldots,x_{i_t})$ 就是 Sobol' 分解中的 $M_{i_1\ldots i_t}(x_{i_1},\ldots,x_{i_t})$，由此可得方差分解：

$$D=\sum_{i_1=1}^n D_{i_1}+\sum_{i_2>i_1}^n D_{i_1i_2}+\cdots+D_{12\ldots n}$$

其中，方差可由基函数系数计算得到：

$$D_{i_1i_2\ldots i_t}=\sum_{\boldsymbol{b}\in I_{i_1\ldots i_t}}a_{\boldsymbol{b}}^2$$

集合 $\mathcal{A}^p$ 中的元素个数随着阶数 p 和维数 n 的增加呈几何级数急剧增加，保留 $\mathcal{A}^p$ 中对应的所有基函数项构建多项式混沌展开所需数据量太大，此外还可能导致过拟合的问题。通常，我们只需保留那些起主要作用的基函数项，其对应的集合 $\mathcal{A}$ 中的元素个数远小于 $\mathcal{A}^p$ 中的元素个数，即构建稀疏多项式混沌展开 (sparse PCE)。

3 模型选择与参数估计

如前所述，构建最优的 PCE 需要解决两个问题，一是选择怎样的 $\mathcal{A}$，即模型选择问题，二是给定 $\mathcal{A}$，如何确定模型的系数向量 $\boldsymbol{a}$，即参数估计问题。下面，先解决参数估计的问题，再解决模型选择的问题。

1. 假设模型：

$$y=\sum_{\boldsymbol{b}\in\mathbb{A}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(\boldsymbol{x})+\varepsilon$$

其中，y 为真实值，$\sum_{\boldsymbol{b}\in\mathbb{A}}a_{\boldsymbol{b}}\Psi_{\boldsymbol{b}}(\boldsymbol{x})$ 为我们所要构建的多项式混沌模型，ε 为真实值与 PCE 拟合值之差，服从一定的概率分布，**而已知 ε 的概率分布事实上就相当于已知似然函数 $L(y|a,\mathcal{A},\boldsymbol{x})$ 或者说概率分布 $p(y|\boldsymbol{a},\mathcal{A},\boldsymbol{x})$，因为省略 $\boldsymbol{x}$ 不会影响我们接下来的推导，为了简便起见我们将 $\boldsymbol{x}$ 省略，分别记为 $L(y|a,\mathcal{A})$ 、$p(y|\boldsymbol{a},\mathcal{A})$，下同。**上式对于变量 $\boldsymbol{x}$ 而言虽然是非线性模型，但是因为 $\Psi(\cdot)$ 的形式已知，若给定 $\boldsymbol{x}$，则 $\Psi(\boldsymbol{x})$ 也已知，将 $\Psi(\boldsymbol{x})$ 视为一个变量，则**上式实际上是一个 y 与 Ψ 的线性模型。**因此，我们要解决的是一个**线性回归问题。**

2. 采样数据：

$$\boldsymbol{y}=\{y_i\}_{i=1}^N,\ \ \boldsymbol{X}=\{x_{i1},x_{i2},\cdots,x_{in}\}_{i=1}^N$$

n 为输入参数分量的个数，N 为样本容量。N 个数据点对应的误差记为 $\boldsymbol{\varepsilon}=\{\varepsilon_i\}_{i=1}^N$，将 N 个数据组合起来：

$$\boldsymbol{y}=\Psi\boldsymbol{a}+\boldsymbol{\varepsilon}$$

其中, $\boldsymbol{a}$ 为系数向量。

3. 假设似然函数：

我们假设 $\varepsilon\sim\mathcal{N}(0,\sigma_\varepsilon^2)$，其中，$\sigma_\varepsilon^2$ 为待定参数。可得 $\boldsymbol{\varepsilon}\sim\mathcal{N}(0,C_M)$，其中，$C_M=\sigma_\varepsilon^2 I_N$，$I_N$ 为 $N\times N$ 的单位矩阵。则 N 个样本点组成的似然函数 $p(\boldsymbol{y}|\boldsymbol{a},C_M,\mathcal{A})\sim\mathcal{N}(\boldsymbol{a},C_M)$：

$$p(\boldsymbol{y}|\boldsymbol{a},C_M,\mathcal{A})=(2\pi)^{-N/2}|C_M|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\Psi\boldsymbol{a})^TC_M^{-1}(\boldsymbol{y}-\Psi\boldsymbol{a})\right)$$

or

$$p(\boldsymbol{y}|\boldsymbol{a},\sigma_\varepsilon^2,\mathcal{A})=(2\pi\sigma_\varepsilon^2)^{-N/2}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})\right)$$

其中，$|\cdot|$ 表示求行列式，其中待定参数有：$\boldsymbol{a}$ 、σ_ε^2 。

注意：

- 若随机变量 X 的样本空间 $\mathbb{X}=\{q_1,q_2,\cdots,q_r\}$，条件概率 $P(Y|X)$ 实际上是由 $P(Y|X=q_1),P(Y|X=q_2),\cdots,P(Y|X=q_r)$ 各个模型组成，各个模型的类型都可以不一样，且只含有变量 Y，不含有变量 X。但若这些模型形式上可以使用 $P(Y|X)$ 统一表示，则条件概率模型 $P(Y|X)$ 不仅含有变量 Y 还含有变量 X，只是这两个变量的地位是不同的，在讨论变量 Y 的概率前，要先对变量 X 进行具体化，即令 $X=q_i$ 。之所以强调上述内容，是因为在下文的推导过程中，我们要牢记，似然函数 $p(\boldsymbol{y}|\boldsymbol{a},\sigma_\varepsilon^2,\mathcal{A})$ 和后验概率 $p(\boldsymbol{a}|\boldsymbol{y},\sigma_\varepsilon^2,\mathcal{A})$ 模型中 $\boldsymbol{y}$ 和 $\boldsymbol{a}$ 均可以看做变量，对于似然函数，$\boldsymbol{a}$ 没有具体化，对于后验概率，$\boldsymbol{y}$ 也没有具体化，不要误认为 $\boldsymbol{y}$ 和 $\boldsymbol{a}$ 已经是一个数值了，只不过在研究似然函数时，我们假设式中 $\boldsymbol{a}$ 已经给定，在研究后验概率时，我们假设式中 $\boldsymbol{y}$ 已经给定。同样地，对待定参数 σ_ε^2 也是如此。总而言之，条件概率 $p(\cdot|\star)$，$\cdot$ 中**是待研究的变量**，$\star$ 中是研究过程中**假设为已知的不改变的"变量"，可以给变量分个等级，$\cdot$ 和 $\star$ 均为变量，但 $\cdot$ 的等级比 $\star$ 高，是变量的变量。**
- **为简便起见，我们并不需要将所有假设为已知的量写入 $\star$ 中，而只需将那些不属于 $\cdot$ 中，但还想研究其变化后的影响的量写入 $\star$ 中。**

4. 假设先验分布：均匀分布，高斯分布，拉普拉斯分布。贝叶斯估计中先验采用均匀分布等价于极大似然估计，采用高斯先验等价于岭回归/ L_2 正则化，采用拉普拉斯先验等价于 LASSO/ L_1 正则化。

我们这里假设系数向量 $\boldsymbol{a}$ 为高斯先验 $p(\boldsymbol{a}|\mathcal{A})=\mathcal{N}(\boldsymbol{a}_0,C_0)$：

$$p(\boldsymbol{a}|\mathcal{A})=(2\pi)^{-P/2}|C_0|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{a}-\boldsymbol{a}_0)^TC_0^{-1}(\boldsymbol{a}-\boldsymbol{a}_0)\right)$$

其中，$\boldsymbol{a}_0$ 和 C_0 为人为设定的参数，是已知的；$P=Card(\mathcal{A})$ 为集合 $\mathcal{A}$ 的基，即多项式系数的个数。

σ_ε^2 的先验概率 $p(\sigma_\varepsilon^2)$ 为均匀分布。

3.1 参数估计

3.1.1 $\boldsymbol{a}$ 的后验估计

系数向量 $\boldsymbol{a}$ 的后验概率由贝叶斯公式可得：

$$p(\boldsymbol{a}|\boldsymbol{y},\sigma_\varepsilon^2,\mathcal{A}) \propto p(\boldsymbol{y}|\boldsymbol{a},\sigma_\varepsilon^2,\mathcal{A})p(\boldsymbol{a}|\sigma_\varepsilon^2,\mathcal{A})$$

实际上，$p(\boldsymbol{a}|\sigma_\varepsilon^2,\mathcal{A})$ 就是 $p(\boldsymbol{a}|\mathcal{A})$，因为 $\boldsymbol{a}$ 的先验概率与 σ_ε^2 无关，因此，此时可认为 σ_ε^2 已给定。因此，有：

$$p(\boldsymbol{a}|\boldsymbol{y},\sigma_\varepsilon^2,\mathcal{A}) \propto$$

$$(2\pi)^{-(N+P)/2}|C_M|^{-1/2}|C_0|^{-1/2}\exp\left(-\frac{1}{2}\left[(\boldsymbol{y}-\Psi\boldsymbol{a})^TC_M^{-1}(\boldsymbol{y}-\Psi\boldsymbol{a})+(\boldsymbol{a}-\boldsymbol{a}_0)^TC_0^{-1}(\boldsymbol{a}-\boldsymbol{a}_0)\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[(\boldsymbol{y}-\Psi\boldsymbol{a})^TC_M^{-1}(\boldsymbol{y}-\Psi\boldsymbol{a})+(\boldsymbol{a}-\boldsymbol{a}_0)^TC_0^{-1}(\boldsymbol{a}-\boldsymbol{a}_0)\right]\right)$$

指数前的系数能够省略是因为其中不含变量 $\boldsymbol{a}$。

现在求 $\boldsymbol{a}$ 的最大后验估计。**以 $\boldsymbol{a}$ 为变量，其他量为已知量**，将上式最大化，就可得到 $\boldsymbol{a}$ 的最大后验估计，记为 $\hat{\boldsymbol{a}}$。一般地，是采用求导并令导数为 0，这里我们也可以采用恒等变形，通过将上式中指数的幂进行配方来求解：

$$(\boldsymbol{y}-\Psi\boldsymbol{a})^TC_M^{-1}(\boldsymbol{y}-\Psi\boldsymbol{a})+(\boldsymbol{a}-\boldsymbol{a}_0)^TC_0^{-1}(\boldsymbol{a}-\boldsymbol{a}_0)=$$

$$\left[\left(\boldsymbol{a}-[\Psi^TC_M^{-1}\Psi+C_0^{-1}]^{-1}[\Psi^TC_M^{-1}\boldsymbol{y}+C_0^{-1}\boldsymbol{a}_0]\right)^T(\Psi^TC_M^{-1}\Psi+C_0^{-1})^{1/2}\right]^2+C$$

上式中 $[\cdot]^2=[\cdot]^T[\cdot]$，$C$ 为常数。可得后验估计表达式：

$$\hat{\boldsymbol{a}}=[\Psi^TC_M^{-1}\Psi+C_0^{-1}]^{-1}[\Psi^TC_M^{-1}\boldsymbol{y}+C_0^{-1}\boldsymbol{a}_0]$$

将采样数据代入，就可计算出 $\hat{\boldsymbol{a}}$，此时 $\hat{\boldsymbol{a}}$ 就被具体化为一个数值了。细心的读者应该注意到了，事实上，目前我们还无法计算 $\hat{\boldsymbol{a}}$，因为其中还含有待估计的量 $C_M=\sigma_\varepsilon^2I_N$，下面我们也会采用同样的方式推导 σ_ε^2 的估计式，并和这里 $\hat{\boldsymbol{a}}$ 的计算式一起组成迭代的计算公式。

可以看到，$\boldsymbol{a}$ 的后验分布也是一个高斯分布 (高斯分布与高斯分布共轭，这里先验分布和似然函数均为高斯分布)，其均值为 $\hat{\boldsymbol{a}}$，则其协方差矩阵 $\hat{C}_{aa}$ 的计算表达式由上面求 $\hat{\boldsymbol{a}}$ 时配方出的公式可知为：

$$\hat{C}_{aa}=(\Psi^TC_M^{-1}\Psi+C_0^{-1})^{-1}$$

这里就可以看到采用配方的方法而不是求导的方法求 $\hat{\boldsymbol{a}}$ 的好处了，因为已知 $\boldsymbol{a}$ 的后验概率为高斯分布，采用配方法求 $\hat{\boldsymbol{a}}$ 时可顺便得到 $\hat{C}_{aa}$，从而可得系数向量 $\boldsymbol{a}$ 的后验概率为：

$$p(\boldsymbol{a}|\boldsymbol{y},\sigma_\varepsilon^2,\mathcal{A})=\mathcal{N}(\hat{\boldsymbol{a}},\hat{C}_{aa})$$

再次强调，若 $\hat{\boldsymbol{a}}$ 和 $\hat{C}_{aa}$ 未被具体化，那就是变量，若已被具体化为具体数值，上式就是确定的概率分布。样本的二重性确实很容易让人混淆，因此在这里反复强调。**这里求后验概率的目的是为了后面计算 BME。**

3.1.2 σ_ε^2 的后验估计

σ_ε^2 的后验概率由贝叶斯公式可得：

$$p(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{a},\mathcal{A}) \propto p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})p(\sigma_\varepsilon^2|\boldsymbol{a},\mathcal{A})$$

和前面类似，$p(\sigma_\varepsilon^2|\boldsymbol{a},\mathcal{A})$ 实际上就是 $p(\sigma_\varepsilon^2)$，因为 σ_ε^2 的先验概率与 $\boldsymbol{a}$ 和 $\mathcal{A}$ 无关，因此，此时可认为 $\boldsymbol{a}$ 和 $\mathcal{A}$ 已给定。前面我们假设 $p(\sigma_\varepsilon^2)$ 为均匀分布，因此：

$$p(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{a},\mathcal{A}) \propto p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})$$

现在求 σ_ε^2 的最大后验估计。**以 σ_ε^2 为变量，其他量为已知量**，将上式最大化，就可得到 σ_ε^2 的最大后验估计，记为 $\hat{\sigma}_\varepsilon^2$。这里采用求导法：

$$\arg\min_{\sigma_\varepsilon^2} p(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{a},\mathcal{A}) \Longleftrightarrow \arg\min_{\sigma_\varepsilon^2} p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})$$

$$-2\ln\left(p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})\right) = N\ln\left(\sigma_\varepsilon^2\right) + \frac{(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})}{\sigma_\varepsilon^2} + C$$

$$-2\frac{d\ln\left(p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})\right)}{d\sigma_\varepsilon^2}\Big|_{\sigma_\varepsilon^2=\hat{\sigma}_\varepsilon^2} = \frac{N}{\hat{\sigma}_\varepsilon^2} - \frac{(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})}{\hat{\sigma}_\varepsilon^4} = 0$$

$$\hat{\sigma}_\varepsilon^2 = \frac{(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})}{N}$$

注意，这里不要对 $\ln p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})$ 求导产生疑惑：根据条件概率的意思， σ_ε^2 为给定的量，为什么还可以对它求导呢？其理解可参见本章开头第三点假设似然函数部分后面注意的内容。其实，在进行极大似然估计的时候也是像这样在对似然函数求导，这里先验分布是均匀分布，因此和使用极大似然估计的结果是一样的。

同样也可以求得 σ_ε^2 的后验分布 (概率密度函数中的归一化常数可以先不用管，等一切结束后再选取使得概率密度函数满足归一性的常数即可)，由贝叶斯公式：

$$p(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{a},\mathcal{A}) \propto p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})p(\sigma_\varepsilon^2|\boldsymbol{a},\mathcal{A}) = (2\pi\sigma_\varepsilon^2)^{-N/2}\exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})\right)$$

以 σ_ε^2 的视角，解读上述形式为正态分布密度函数的式子 (即以 σ_ε^2 为变量，其他量为已知量)：

$$p(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{a},\mathcal{A}) \propto (\sigma_\varepsilon^{-2})^{k-1}\exp\left(-\frac{\sigma_\varepsilon^{-2}}{\theta}\right) = \Gamma\left(\frac{N+2}{2}, \frac{2}{(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})}\right)$$

3.1.3 $\hat{\boldsymbol{a}}$ 和 $\hat{\sigma}_\varepsilon^2$ 的迭代计算

前面我们分别对 $\boldsymbol{a}$ 和 σ_ε^2 进行最大后验估计，在各自最大后验估计的过程中，它们所使用的似然函数相同，均为 $p(\boldsymbol{y}|\sigma_\varepsilon^2,\boldsymbol{a},\mathcal{A})$，且都假设对方已经给定，由此得到了如下两个"你中有我，我中有你"的估计式：

$$\begin{aligned}\hat{\boldsymbol{a}} &= [\Psi^T C_M^{-1}\Psi + C_0^{-1}]^{-1}[\Psi^T C_M^{-1}\boldsymbol{y} + C_0^{-1}\boldsymbol{a}_0],\ \text{with } C_M = \sigma_\varepsilon^2 I_N \\ \text{or} &= \hat{C}_{aa}[\Psi^T C_M^{-1}\boldsymbol{y} + C_0^{-1}\boldsymbol{a}_0],\ \text{with } \hat{C}_{aa} = (\Psi^T C_M^{-1}\Psi + C_0^{-1})^{-1} \\ \text{or} &= [\sigma_\varepsilon^{-2}\Psi^T\Psi + C_0^{-1}]^{-1}[\sigma_\varepsilon^{-2}\Psi^T\boldsymbol{y} + C_0^{-1}\boldsymbol{a}_0]\end{aligned}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{(\boldsymbol{y}-\Psi\boldsymbol{a})^T(\boldsymbol{y}-\Psi\boldsymbol{a})}{N}$$

其中，除了 $\boldsymbol{a}$ 和 σ_ε^2，其他的量或是由样本数据给定如 $\boldsymbol{y}$、Ψ、N，或是事先人为给定如 C_0、$\boldsymbol{a}_0$。取 $\hat{\boldsymbol{a}}$ 计算公式中的 σ_ε^2 为 $\hat{\sigma}_\varepsilon^2$，取 $\hat{\sigma}_\varepsilon^2$ 计算公式中的 $\boldsymbol{a}$ 为 $\hat{\boldsymbol{a}}$，则可得计算公式：

$$\hat{\boldsymbol{a}} = [\hat{\sigma}_\varepsilon^{-2}\Psi^T\Psi + C_0^{-1}]^{-1}[\hat{\sigma}_\varepsilon^{-2}\Psi^T\boldsymbol{y} + C_0^{-1}\boldsymbol{a}_0]$$

$$\hat{\sigma}_\varepsilon^2 = \frac{(\boldsymbol{y}-\Psi\hat{\boldsymbol{a}})^T(\boldsymbol{y}-\Psi\hat{\boldsymbol{a}})}{N}$$

联立这两个方程，我们发现无法解析地求解，则给定一个初值采用迭代的方法求解。事实上，对于这种情况，更稳妥的做法是选择多个初始值进行迭代，若得到多个收敛点，则选择其中使后验概率最大的那个点作为参数的最大后验估计，但因为本问题是一个凸优化问题，只存在一个解，因此只用选取一个初始值进行迭代。这两个计算公式和论文中的三个计算公式是等价的。

3.1.4 似然函数的计算

基于最开始的假设，似然函数 $p(\boldsymbol{y}|\boldsymbol{a},C_M,\mathcal{A}) \sim \mathcal{N}(\boldsymbol{a},C_M)$：

$$p(\boldsymbol{y}|\boldsymbol{a},C_M,\mathcal{A}) = (2\pi)^{-N/2}|C_M|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\Psi\boldsymbol{a})^T C_M^{-1}(\boldsymbol{y}-\Psi\boldsymbol{a})\right)$$

其中，$C_M = \sigma_\varepsilon^2 I_N$，$\boldsymbol{a}$ 和 σ_ε^2 为待定参数。现在，已经计算出 $\boldsymbol{a}$ 和 σ_ε^2 的后验估计 $\hat{\boldsymbol{a}}$ 和 $\hat{\sigma}_\varepsilon^2$，代入即可得似然函数 $p(\boldsymbol{y}|\hat{\boldsymbol{a}},\hat{C}_M,\mathcal{A}) \sim \mathcal{N}(\hat{\boldsymbol{a}},\hat{C}_M)$：

$$p(\boldsymbol{y}|\hat{\boldsymbol{a}},\hat{C}_M,\mathcal{A}) = (2\pi)^{-N/2}|\hat{C}_M|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\Psi\hat{\boldsymbol{a}})^T \hat{C}_M^{-1}(\boldsymbol{y}-\Psi\hat{\boldsymbol{a}})\right)$$

其中，$\hat{C}_M = \hat{\sigma}_\varepsilon^2 I_N$ 。

3.2 模型选择

3.2.1 KIC (Kashyap information criterion) 的推导

本小节的推导不基于任何先验假设和似然假设，具有一般性。

Bayesian model selection 选择使模型的后验概率 $p(\mathcal{A}|\boldsymbol{y})$ 最大的模型 $\mathcal{A}$：

$$p(\mathcal{A}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mathcal{A})p(\mathcal{A})$$

模型的先验概率设为均布分布，则 $p(\mathcal{A}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mathcal{A})$，而 $p(\boldsymbol{y}|\mathcal{A})$：

$$p(\boldsymbol{y}|\mathcal{A}) = \int_{\mathbb{R}^P} p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A})d\boldsymbol{a} = \int_{\mathbb{R}^P} p(\boldsymbol{y}|\mathcal{A},\boldsymbol{a})p(\boldsymbol{a}|\mathcal{A})d\boldsymbol{a}$$

类比参数估计，上式相当于在进行“极大似然估计”，或者说是先验概率为均匀分布的“最大后验估计”。

接下来，我们采用近似的方法计算上述积分。以 $\boldsymbol{a}$ 为变量，将被积函数 $p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A})$ 取对数后在其 MAP 处进行泰勒展开，忽略三阶及三阶以上的项 (**当参数的后验分布是高斯分布时，KIC@MAP 就是BME的精确解，因为此时三阶及三阶以上导数为** 0)：

$$\ln p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A}) \approx \ln p(\boldsymbol{y},\hat{\boldsymbol{a}}|\mathcal{A}) + (\boldsymbol{a}-\hat{\boldsymbol{a}})^T \left[\frac{d\ln p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A})}{d\boldsymbol{a}}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right] + \frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T \left[\frac{d^2\ln p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A})}{d\boldsymbol{a}^2}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right](\boldsymbol{a}-\hat{\boldsymbol{a}})$$

其中，

- 因为 $p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A}) = p(\boldsymbol{a}|\boldsymbol{y},\mathcal{A})p(\boldsymbol{y}|\mathcal{A})$，而 $p(\boldsymbol{y}|\mathcal{A})$ 不含 $\boldsymbol{a}$，结合 MAP 的定义可知一阶导数：
$$\left[\frac{d\ln p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A})}{d\boldsymbol{a}}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right] = \left[\frac{d\ln p(\boldsymbol{a}|\boldsymbol{y},\mathcal{A})}{d\boldsymbol{a}}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right] = \boldsymbol{0}$$
- 同时，为表述方便，记 MAP 处的负二阶导为 $\hat{\Sigma}^{-1}$：
$$\hat{\Sigma}^{-1} = -\left[\frac{d^2\ln p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A})}{d\boldsymbol{a}^2}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right] = -\left[\frac{d^2\ln p(\boldsymbol{a}|\boldsymbol{y},\mathcal{A})}{d\boldsymbol{a}^2}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right]$$

则：

$$\ln p(\boldsymbol{y},\boldsymbol{a}|\mathcal{A}) \approx \ln p(\boldsymbol{y},\hat{\boldsymbol{a}}|\mathcal{A}) - \frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T\hat{\Sigma}^{-1}(\boldsymbol{a}-\hat{\boldsymbol{a}})$$

代入到积分公式中可得：

$$p(\boldsymbol{y}|\mathcal{A}) \approx p(\boldsymbol{y},\hat{\boldsymbol{a}}|\mathcal{A})\int_{\mathbb{R}^P}\exp\left(-\frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T\hat{\Sigma}^{-1}(\boldsymbol{a}-\hat{\boldsymbol{a}})\right)d\boldsymbol{a}$$

而：

$$\int_{\mathbb{R}^P}\exp\left(-\frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T\hat{\Sigma}^{-1}(\boldsymbol{a}-\hat{\boldsymbol{a}})\right)d\boldsymbol{a} =$$
$$(2\pi)^{P/2}|\hat{\Sigma}|^{1/2}\int_{\mathbb{R}^P}\frac{1}{\sqrt{(2\pi)^P|\hat{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T\hat{\Sigma}^{-1}(\boldsymbol{a}-\hat{\boldsymbol{a}})\right)d\boldsymbol{a} = (2\pi)^{P/2}|\hat{\Sigma}|^{1/2}$$

上式中的积分恰为正态分布密度函数的积分，因此结果为 1，从而可得：

$$p(\boldsymbol{y}|\mathcal{A}) \approx p(\boldsymbol{y},\hat{\boldsymbol{a}}|\mathcal{A})(2\pi)^{P/2}|\hat{\Sigma}|^{1/2} = p(\boldsymbol{y}|\hat{\boldsymbol{a}},\mathcal{A})p(\hat{\boldsymbol{a}}|\mathcal{A})(2\pi)^{P/2}|\hat{\Sigma}|^{1/2}$$

则 KIC 定义下：

$$-2\ln p(\boldsymbol{y}|\mathcal{A}) \approx -2\ln p(\boldsymbol{y}|\hat{\boldsymbol{a}},\mathcal{A}) - 2\ln p(\hat{\boldsymbol{a}}|\mathcal{A}) - P\ln(2\pi) - \ln|\hat{\Sigma}| \triangleq KIC$$

3.2.2 BSPCE 的 KIC

上小节推导 KIC 时的似然函数 $p(\boldsymbol{y}|\mathcal{A},\boldsymbol{a})$ 和参数先验分布 $p(\boldsymbol{a}|\mathcal{A})$，对于本问题就是 $p(\boldsymbol{y}|\mathcal{A},\boldsymbol{a},\hat{\sigma}_\varepsilon^2)$ 和 $p(\boldsymbol{a}|\mathcal{A})$。

对于本问题而言，后验分布 $p(\boldsymbol{a}|\boldsymbol{y},\hat{\sigma}_\varepsilon^2,\mathcal{A})\sim\mathcal{N}(\hat{\boldsymbol{a}},\hat{C}_{aa})$，为正态分布，则：

$$\hat{\Sigma}^{-1}=-\left[\frac{d^2\ln p(\boldsymbol{a}|\boldsymbol{y},\hat{\sigma}_\varepsilon^2,\mathcal{A})}{d\boldsymbol{a}^2}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right]$$

$$=-\left[\frac{d^2\ln\left((2\pi)^{-P/2}|\hat{C}_{aa}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T\hat{C}_{aa}^{-1}(\boldsymbol{a}-\hat{\boldsymbol{a}})\right)\right)}{d\boldsymbol{a}^2}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right]$$

$$=-\left[\frac{d^2\left(-\frac{1}{2}(\boldsymbol{a}-\hat{\boldsymbol{a}})^T\hat{C}_{aa}^{-1}(\boldsymbol{a}-\hat{\boldsymbol{a}})\right)}{d\boldsymbol{a}^2}\Big|_{\boldsymbol{a}=\hat{\boldsymbol{a}}}\right]$$

$$=\hat{C}_{aa}^{-1},\ (\text{求导公式：}\frac{\partial\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial\boldsymbol{x}}=(\boldsymbol{A}+\boldsymbol{A}^T)\boldsymbol{x})$$

即对于正态后验，$\hat{\Sigma}=\hat{C}_{aa}$。由此可得本问题的模型选择指标的计算公式：

$$KIC=-2\ln p(\boldsymbol{y}|\hat{\boldsymbol{a}},\hat{\sigma}_\varepsilon^2,\mathcal{A})-2\ln p(\hat{\boldsymbol{a}}|\mathcal{A})-P\ln(2\pi)-\ln|\hat{C}_{aa}|$$

其中，$\hat{\boldsymbol{a}}$ 和 $\hat{C}_{aa}$ 由参数估计中的迭代公式计算得到，参数后验估计 $p(\hat{\boldsymbol{a}}|\mathcal{A})\sim\mathcal{N}(\hat{\boldsymbol{a}},\hat{C}_{aa})$，似然函数 $p(\boldsymbol{y}|\hat{\boldsymbol{a}},\hat{C}_M,\mathcal{A})\sim\mathcal{N}(\hat{\boldsymbol{a}},\hat{C}_M)$。

至此，BSPCE 的参数估计和模型选择的推导全部结束。

3.3 模型选择与参数估计的辩证关系

- 贝叶斯理论要解决的两个问题，首先是模型选择，然后是模型参数的估计。模型选择需要评估 BME，而 BME 的近似计算需要用到参数估计的结果。
- 先验概率为均匀分布时的最大后验估计实际上就是极大似然估计。模型选择是可以看做是更高层次的参数估计，因为“模型选择”和“参数估计”的界限本身就是人为给定的。比如，构建 PCE 时模型选择就是确定模型包含哪些多项式基函数，也就是对索引 $\boldsymbol{b}$ 进行估计；而 PCE 模型的参数估计就是在基函数已经确定的情况下，在各种具有不同多项式系数 $\boldsymbol{a}$ 的 PCE 模型中挑选出最优的模型。从这个角度看，基于 BME 进行模型选择也相当于在进行极大似然估计，或者说模型先验为均匀分布的最大后验估计。
- 贝叶斯模型选择实际上是在评估平均似然：

$$p(\mathcal{A}|\boldsymbol{y})\propto p(\boldsymbol{y}|\mathcal{A})p(\mathcal{A})\propto p(\boldsymbol{y}|\mathcal{A})$$

$$p(\boldsymbol{y}|\mathcal{A})=\int_{\mathbb{R}^P}p(\boldsymbol{y}|\mathcal{A},\boldsymbol{a})p(\boldsymbol{a}|\mathcal{A})d\boldsymbol{a}$$

其中 $p(\boldsymbol{y}|\mathcal{A},\boldsymbol{a})$ 为参数 $\boldsymbol{a}$ 给定后的似然函数，$p(\boldsymbol{a}|\mathcal{A})$ 为参数 $\boldsymbol{a}$ 的先验概率，则**上述积分就是在计算平均似然，BME 评估的是模型在各种参数设置下拟合样本数据的平均能力。**由此，引申出一个问题；模型 $\mathcal{A}_1$ 的平均似然比模型 $\mathcal{A}_2$ 的大，但 $\mathcal{A}_1$ 和其参数的后验估计 $\hat{\boldsymbol{a}}_1$ 构成的模型比 $\mathcal{A}_2$ 和其参数的后验估计 $\hat{\boldsymbol{a}}_2$ 构成的模型对样本数据拟合得更好，即：

$$p(\boldsymbol{y}|\mathcal{A}_1)=\int_{\mathbb{R}^P}p(\boldsymbol{y}|\mathcal{A}_1,\boldsymbol{a})p(\boldsymbol{a}|\mathcal{A}_1)d\boldsymbol{a}>p(\boldsymbol{y}|\mathcal{A}_2)=\int_{\mathbb{R}^P}p(\boldsymbol{y}|\mathcal{A}_2,\boldsymbol{a})p(\boldsymbol{a}|\mathcal{A}_2)d\boldsymbol{a}$$

$$\text{while}$$

$$p(\boldsymbol{y}|\mathcal{A}_1,\hat{\boldsymbol{a}}_1)<p(\boldsymbol{y}|\mathcal{A}_2,\hat{\boldsymbol{a}}_2)$$

那么，基于 BME 不就选择不到对样本数据拟合得最好的模型了么，贝叶斯模型选择还合理么？答案当然是合理的。若同时考虑所有模型 $\mathcal{A}$ 和所有参数 $\boldsymbol{a}$，选择其中对样本数据拟合得最好的模型，实际上就在进行极大似然估计；而**先基于 BME 选择模型 $\mathcal{A}$，再对模型 $\mathcal{A}$ 的参数 $\boldsymbol{a}$ 进行估计，就相当于在进行正则化，综合考虑了对数据的拟合程度和模型的复杂程度**，因此基于 BME 选择的模型不是对样本数据拟合得最好的是很正常的。**这在某种程度上可以看出贝叶斯模型选择是遵循奥卡姆剃刀 (Occam's razor) 原理的。除此之外，可在贝叶斯模型选择的基础上，进一步采用交叉验证等模型选择方法对构建出来的模型进行评估。**

- 可以这样理解：参数估计偏向于更好地拟合训练数据，而模型选择还考虑了模型复杂度、过拟合、以及模型在预测数据上的表现，即泛化能力等方面的问题。

4 BSPCE 的构建算法及展望

首先说明下实际计算过程中对各个数据的处理：

N：样本容量。

$\boldsymbol{y}$：对采样得到的 N 个模型响应进行标准化，然后组成响应向量 $\boldsymbol{y}$。因此，响应样本数据的均值为 0，方差为 1。**事实上，可以看到 PCE 的理论没有要求对模型响应进行标准化，可不做任何处理。这里这样处理是为了方便假定先验分布。**

$\boldsymbol{X}$：模型输入需为 $[0,1]$ 上的均匀分布，否则还需进行相应的转化，这是我们所用 PCE 基函数的要求。

Ψ：由模型输入样本矩阵 $\boldsymbol{X}$ (不说输入向量是因为每个输入可能含有分量)和基函数计算确定。

$\boldsymbol{a}_0$：系数向量先验分布的均值，因为 $\boldsymbol{y}$ 是标准化后的数据，$\Psi\boldsymbol{a}_0$ 是对 $\boldsymbol{y}$ 的近似，因此，不妨令 $\boldsymbol{a}_0=\boldsymbol{0}$

C_0：系数向量先验分布的协方差矩阵：

$$\boldsymbol{C_{aa}}=\begin{bmatrix}\delta_{\boldsymbol{b}_1}^2 & 0 & \cdots & \cdots\\ 0 & \ddots & 0 & \cdots\\ \vdots & 0 & \ddots & 0\\ 0 & \cdots & 0 & \delta_{\boldsymbol{b}_P}^2\end{bmatrix}$$

$$\text{with }\ \delta_{\boldsymbol{b}_i}^2=(p_{\boldsymbol{b}_i}+h_{\boldsymbol{b}_i}-1)h_{\boldsymbol{b}_i}^2$$

$$\text{where for any index }\boldsymbol{b},\ \ p_{\boldsymbol{b}}\equiv|\boldsymbol{b}|=\sum_{i=1}^{n}b_i,\ \ h_{\boldsymbol{b}}\equiv\sum_{i=1}^{n}\mathbf{1}_{b_i>0}$$

其中，$p_{\boldsymbol{b}_i}$ 和 $h_{\boldsymbol{b}_i}$ 分别对应 $\mathcal{A}$ 中第 i 项基函数的阶数和交互度，$\delta_{\boldsymbol{b}_i}^2$ 对应 $\mathcal{A}$ 中第 i 项基函数的方差。可见，该先验假设对总阶数较高且交互度较高的基函数项赋予了较大的方差，而对总阶数较低且交互度较低的基函数项赋予了较小的方差。**这样设置协方差矩阵给出的偏好是**：低阶低交互度基函数的系数偏向于要么很大要么很小，取到 0 的可能性也比较高；而高阶高交互度基函数的系数偏向于在 0 附近，但取到 0 的可能性比低阶低交互度的情况低。可参照 L_1 和 L_2 正则化的区别进行理解。

4.1 BSPCE 的构建算法

Bayesian Sparse PCE 的构建算法如下：

1. 初始化

- 数据预处理：将输入映射为 $[0,1]$ 上的均匀分布，将输出数据进行标准化；
- 设定初始多项式最高阶 p 及最大交互度 h，一般可令 $(p=2,h=1)$ 或 $(p=4,h=2)$；
- 构造索引集合：$\mathcal{A}^{p,h}=\{\boldsymbol{a}\in\mathbb{N}^n:p_{\boldsymbol{a}}\leqslant p,h_{\boldsymbol{a}}\leqslant h\}$；
- 构造基函数向量：$\Psi=(\Psi_1,\Psi_2,\ldots,\Psi_P)$，$P=card(\mathcal{A}^{p,h})$。

2. 根据相关系数对基函数排序

- 基于基函数向量 Ψ 计算每个基函数与模型响应的皮尔逊相关系数：

$$r_j=\frac{\mathbb{COV}[\boldsymbol{y},\Psi_j(\boldsymbol{x})]}{\sqrt{\mathbb{V}[\boldsymbol{y}]\mathbb{V}[\Psi_j(\boldsymbol{x})]}}$$

- 对基函数向量 Ψ 进行排序得到：

$$\hat{\Psi}=(\hat{\Psi}_1,\ldots,\hat{\Psi}_j,\hat{\Psi}_{j+1}\ldots,\hat{\Psi}_P)$$

其中，$\hat{r}_j^2\geqslant\hat{r}_{j+1}^2$。

3. 根据偏相关系数对基函数再排序

- 基于基函数向量 $\hat{\Psi}$ 计算每个基函数与模型响应的偏相关系数：

$$\hat{r}_{j|1,\ldots,j-1}=\frac{\mathbb{COV}[\boldsymbol{y},\hat{\Psi}_j(\boldsymbol{x})|\hat{\Psi}_1(\boldsymbol{x}),\ldots,\hat{\Psi}_{j-1}(\boldsymbol{x})]}{\sqrt{\mathbb{V}[\boldsymbol{y}|\hat{\Psi}_1(\boldsymbol{x}),\ldots,\hat{\Psi}_{j-1}(\boldsymbol{x})]\mathbb{V}[\Psi_j(\boldsymbol{x})|\hat{\Psi}_1(\boldsymbol{x}),\ldots,\hat{\Psi}_{j-1}(\boldsymbol{x})]}}$$

- 对基函数向量 $\hat{\Psi}$ 进行再排序以消除伪相关性：

$$\tilde{\Psi}=(\tilde{\Psi}_1,\ldots,\tilde{\Psi}_j,\tilde{\Psi}_{j+1}\ldots,\tilde{\Psi}_P)$$

其中，$\tilde{r}_{j|1,\ldots,j-1}^2\geqslant\tilde{r}_{j+1|1,\ldots,j}^2$。

4. 对基函数进行识别/模型选择

- 令基函数向量 $\Psi_{\mathcal{A}^{opt}}=\tilde{\Psi}_1$，$KIC_1=+\infty$，$k=2$；
- 基于基函数向量 $\Psi_{\mathcal{A}_k}=(\Psi_{\mathcal{A}^{opt}},\tilde{\Psi}_k)$，进行参数估计，构建多项式混沌展开 $M_{\mathcal{A}_k}$，计算并比较 $M_{\mathcal{A}_k}$ 的 KIC_k 与 KIC_{k-1} 的大小，判断 $\tilde{\Psi}_k$ 是否加入最优基函数向量 $\Psi_{\mathcal{A}^{opt}}$；
- 置 $k=k+1$，重复上述判定直到 $k=P$，得 $M_{\mathcal{A}^{opt}}$。

5. 扩充基函数向量

- 若 $M_{\mathcal{A}^{opt}}$ 中存在总阶数超过 $p-1$ 或交互项的阶达到 h 的高阶项，则令 $p=p+2$ 或 $h=h+1$。接着，对已识别出的多项式基函数向量 $\Psi_{\mathcal{A}^{opt}}$ 中的高阶项进行升阶操作，在总阶数不超过 p 且交互项的阶不超过 h 的条件下构造出对应的更高阶的多项式基函数，并和 $\Psi_{\mathcal{A}^{opt}}$ 一起，组成更新过后的多项式基函数向量 Ψ，接着重复2-5步；
- 若不存在，则输出最优的多项式混沌展开 $M_{\mathcal{A}^{opt}}$；
- 若模型残差 ε 的方差除以总方差即 $\hat{\sigma}_\varepsilon^2/D$ 小于 5%，则表明构造出的SPCE对样本数据的拟合程度较高；否则增大样本数，重复本算法。

我们补充说明如下：

1. **总方差 D 的计算与上一章 PCE 中给出的稍有不同**，根据我们对模型的假设总方差：$D=\sum_{\boldsymbol{b}\in\mathcal{A}^{opt}}a_{\boldsymbol{b}}^2+\hat{\sigma}_\varepsilon^2$，除了系数的平方项外还需加上残差 ε 的方差 $\hat{\sigma}_\varepsilon^2$。
2. **如何求 Sobol' 指数的置信区间。**显然，我们是在贝叶斯的框架下进行的分析，因此，这里的置信区间取贝叶斯派的含义而不是频率派的含义。求 Sobol' 指数 95% 的置信区间的具体操作是：首先对 $\boldsymbol{a}$ 按照其后验分布在整个域内进行采样 (正态分布对应拉丁超立方取样，共取了 $N_{\boldsymbol{a}}=100,000$ 个样本)，再计算每个样本点对应的 Sobol' 指数并对 Sobol' 指数进行排序，去掉最小的 $2.5\%N_{\boldsymbol{a}}$ 及最大的 $2.5\%N_{\boldsymbol{a}}$ 个 Sobol' 指数，就可由剩下的样本组成 Sobol' 指数 95% 的置信区间。事实上，每个样本点都等概率的出现 (注意，不是每种可能取值都等概率地出现，不同样本点的取值可能相同，根据每个样本点 Sobol' 指数的值进行排序可以想象为在用柱状图画 Sobol' 指数的概率分布)，排序后去掉首尾 5% 个样本点，剩下的样本点所分布的范围就是 Sobol' 指数 95% 的置信区间。
3. **问题**：为何我们构建的多项式混沌展开是稀疏多项式混沌张开，**其稀疏性从哪里得到保证**?

- 首先是多项式系数 $\boldsymbol{a}$ 的参数估计，其先验分布采用高斯先验，相当于 L_2 正则化。虽然采用拉普拉斯先验，即 L_1 正则化更容易使 $\boldsymbol{a}$ 中出现为 0 的分量，但相比于不引入先验分布，即采用极大似然估计，L_2 正则化还是有助于多项式系数 $\boldsymbol{a}$ 中 0 的出现的。但仅凭正则化的参数估计还不能保证多项式混沌展开的稀疏性，更多的是从构建算法上来保证。此外，基于 KIC 进行模型选择也有助于稀疏解的产生。
- 我们注意到，本节所述算法在第5步扩充基函数向量时，仅仅只是基于原基函数向量中的高阶项进行衍生，得到新的较少的更高阶的多项式基函数，再加入原基函数向量中得到新的基函数向量，因此可以保证所构建得多项式混沌展开的稀疏性。

4.2 BSPCE 算法的展望

算法有待进一步研究的方面有：

1. 理论上来说，采用拉普拉斯似然和拉普拉斯先验的结果会更加好，但计算比较复杂。
2. 这里我们采用的是 KIC 进行模型选择，我们还可以采用其他的指标，比如 BIC 、AIC 进行模型选择。其中，AIC 在小样本时表现较好，也是一个研究方向。
3. 模型的最终评估方法有待改进。 $\hat{\sigma}_{\varepsilon}^{2} / D \leqslant 5\%$ 表明模型几乎捕捉到了样本所有的方差，但这也只能说明构建出来的模型对训练数据拟合得较好。事实上，更好的做法是采用交叉验证 (比如，留一法) 对模型的好坏 (也就是泛化性能) 进行判断。
4. 算法第 5 步基函数的扩充策略可以作进一步的改进。比如，像遗传算法一样采用启发式的方法扩充基函数，通过加入一些随机性，使算法更加健壮。
5. 集成学习。可基于 BMA (Bayesian model averaging) 集成多个 BSPCE 模型，Prof. Mara 已经在做模型的集成，但思路与 BME 有所区别，但他说确实可以基于 BME 进行集成，这也是一个研究方向。

5 文献阅读之 KIC、BIC、AIC

文献[11]是 KIC 的起源，但文章解决的问题是如何构建决策函数。其中，文章第三节前半部分的内容就类似于采用 $0-1$ 损失+期望风险最小化推导贝叶斯决策论，第三节后半部分的内容则给出了损失函数取更一般时的最优决策函数。只不过，在这个过程中需要用到模型的后验概率，其推导在附录中给出，虽然里面并没有显示地提出 KIC 的概念，但可以看到其推导过程。

由文献[9]可知，BMA (Bayesian model averaging) 实际上就是一种集成学习方法，将各种可能模型加权平均构成最终模型，各模型的权重就是模型的后验概率。而计算模型的后验概率，就等价于计算 BME (Bayesian model evidence)。因此，BME 起源于 BMA。显然，若只挑选一个模型，则BME 可被用作模型选择的指标。

文献[9]提到对参数进行贝叶斯估计时参数的先验概率的选择问题，存在两个学派，一种是**主观贝叶斯 (subject Bayesianism)**，一种是**客观贝叶斯 (object Bayesianism)**。主观贝叶斯认为先验概率的选取应该表现出选择人对参数的一些先验观念，而客观贝叶斯则认为选取的先验概率要在某种意义是不提供任何信息的 (noninformative)，一种常用的**无信息先验**是Jeffrey's prior。无信息先验常常不是一个正确的概率分布，因为 $\int p(\theta) d\theta = \infty$ (比如，均匀分布 $p(\theta) = 1$)，此时，对任意正数 c ， $cp(\theta)$ 也是一个无信息先验，但采用 $p(\theta)$ 和 $cp(\theta)$ 计算，得到的后验概率是相同的。无信息先验中未定义常数 c 对参数估计没有影响，因为参数估计只涉及一个模型，但会给模型选择造成困扰。因为对两个不同的模型计算它们的 BME 时，可分别采用无信息先验 $c_1 p(\theta)$ 和 $c_2 p(\theta)$ ，因为常数 c_1 和 c_2 是任意给定的，这样根据它们的 BME 就无法从中挑选出更优模型 (有人提出了the theory of intrinsic Bayes factors 来解决这一问题)。无信息先验可见综述 Kass and Wasserman (1996)。文章最后还提到了一种叫 robust Bayesian inference 的理论，其中贝叶斯推断是基于先验分布集而不是单个先验分布，这为主观贝叶斯和客观贝叶斯搭建了一个桥梁。

文献[9]比较了 AIC 和 BIC，指出 AIC 不是基于 BME 而是基于 KL 散度推得的，AIC 在小样本情况下更加适用，而 BIC 在样本数量趋于无穷时会收敛于选择真实模型 (若真实模型存在于备选模型中) 。文献[9]还讨论了若备选模型中不存在真实模型的问题，并指出在这种情况下会选择与真实模型最相近的模型。 BME 还遵循奥卡姆剃刀 (Occam's razor) 原则。

文献[9]给出的一个回归的问题就是我们构建 BSPCE 所面临的问题。

文献[8]深度好文。

文献[8]给出了BME的两种表达式，一种是常用的基于积分的表达形式，一种是非积分的表达形式，并系统阐述和比较了求 BME 的两种解析法和九种非解析的近似计算方法，包括：1) 两种解析法，分别基于BME的积分表达式和非积分表达式，能精确计算 BME，但均需基于很强的假设条件；2) 四种数值计算法，计算开销大；3) 通过信息准则近似估计BME，五种：AIC、AICc、BIC、KIC@MLE、KIC@MAP (事实上AIC不属于BME的理论框架内，它是由 KL 散度推得，只不过最后的形式和BIC、KIC等基于 BME 理论的指标比较相似，但也常被用于 BMA)。各种方法的假设及相关结论可参见文献中的 Table 4 。

文献[8]指出贝叶斯模型选择理论隐式地遵循了奥卡姆剃刀 (Occam's razor) 原则 (a compromise between model complexity and goodness of fit (also known as the bias-variance trade-off))，并进行了解释。

KIC基于拉普拉斯近似/拉普拉斯方法 (Laplace method [De Beuijn,1961])，拉普拉斯近似是一种近似求积分的方法，其思想是将原函数的积分近似为一个更简单的近似函数的积分。其中，原函数在某个子域内高度集中，而在剩余区域的积分可忽略不计，从而可以寻找一个局部近似的函数(比如泰勒展开)，来替代原函数在该子域内进行积分。

KIC@MAP 和 KIC@MLE： KIC的推导是基于 MAP 估计量进行的，但若参数先验的影响很小，KIC公式中的 MAP 估计量可用 MLE 估计量近似，因为相对而言，MLE估计量更好求解。但还是推荐使用 KIC@MAP，因为其在理论上是一致的，此外根据文中所说，当参数的后验分布也是高斯分布时，KIC@MAP 就是BME的精确解。

KIC@MAP Σ 和 C_{uu} 近似，当参数后验分布为高斯分布时二者相等。

BIC 只是截取了 KIC@MLE 的一部分，理论支撑不足，随着样本数的增加 BIC 和 KIC@MLE 渐进相等。当数据集无穷大时，基于 BIC 和 KIC(@MAP、@MLE)，真实模型一定会被选中 (若真实模型在备选的模型中)。虽然 BIC 比 KIC 更短，但有时在选择模型时更具解释性，而且因为不需要估计海瑟矩阵，其计算量更小。BIC 的一种计算公式如下：

$$BIC = -2\ln p(\boldsymbol{y}|\tilde{\boldsymbol{a}}, \mathcal{A}) + P\ln N$$

其中，$\tilde{\boldsymbol{a}}$ 为极大似然估计量，P 为参数个数，N 为样本容量。

AIC 是从计算 KL 散度或者更确切的说是交叉熵中推得 (事实上只计算交叉熵只能得到 AIC 中的似然项，也就是极大似然估计)，并使用参数个数 P 进行修正，其计算需要用到 MLE 估计量 $\tilde{\boldsymbol{a}}$，采用类似 KIC 中的标记方法即 AIC@MLE：

$$AIC = -2\ln p(\boldsymbol{y}|\tilde{\boldsymbol{a}}, \mathcal{A}) + 2P$$

AICc 是在 AIC 的基础上再加了一个修正项，在小样本时更推荐使用 AICc，随着样本数的增加，AICc 趋于 AIC 。但 AIC 和 AICc 在样本数趋于无穷时均无法保证真实模型一定被选中，因为随着样本数据的增加，AIC(c) 选择的模型的复杂性将会增加，有可能超出真实模型的复杂性 (若真实模型在备选的模型中)。

文献[8]： KIC usually outperforms the BIC and the AIC especially when the model is linear and the error is Gaussian. (线性模型，残差为高斯分布时，使用 KIC 更好)

文献[12]：BIC 的推导，简单易懂。

文献[13]引用次数 6871，没怎么看，可作为 AIC 的参考文献。

文献[10]：KIC@MLE，没有看。

附录 - 概率回归

目标是构建拟合模型 $y_{model} = f(x;\theta)$，其中，θ 是模型的待定参数。则 $y_{real} = f(x;\theta) + \varepsilon$ (为了简便，之后记 y_{real} 为 y)，测量误差和模型误差都打包在随机噪声项 ε 中，并假定其服从一定的概率分布 $\varepsilon \sim p_\varepsilon(\varepsilon)$，其中 $p_\varepsilon(\cdot)$ 表示某个具体的概率密度函数。则 $y - f(x;\theta) \sim p_\varepsilon(y - f(x;\theta))$，也就是 $p(y|x,\theta) = p_\varepsilon(y - f(x;\theta))$，表示在参数 θ 已知的情况下，给定 x 后，y 的概率。可以给变量分个等级，y 的等级比 x 和 θ 的高，x 和 θ 是变量，而 y 是变量的变量。

- 若概率模型使用高斯分布，$\varepsilon \sim \mathcal{N}(0, \sigma^2)$，则

$$p(y|x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y - f(x;\theta))^2}{2\sigma^2}\right)$$

取负对数损失：

$$-\ln p(y|x,\theta) = \frac{(y - f(x;\theta))^2}{2\sigma^2} + C$$

其中，C 为常数。可见，**最小化均方误差等价于对高斯模型进行极大似然估计。**

- 若概率模型使用拉普拉斯分布，$\varepsilon \sim \mathcal{L}aplace(0, \sigma)$，$\sigma > 0$：

$$p(y|x,\theta)=\frac{1}{2\sigma}\exp\left(-\frac{|y-f(x;\theta)|}{\sigma}\right)$$

取负对数损失：

$$-\ln p(y|x,\theta)=\frac{|y-f(x;\theta)|}{\sigma}+C$$

其中，C 为常数。可见，**最小化绝对误差等价于对拉普拉斯模型进行极大似然估计。**

参考文献

1. Global sensitivity analysis using polynomial chaos expansions - Bruno Sudret - 2008
2. Sensitivity estimates for nonlinear mathematical models - I. M. Sobol'- 1993
3. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates - I. M. Sobol'- 2001
4. Physical systems with random uncertainties: Chaos representations with arbitrary probability measure - Christian Soize, R. Ghanem - 2012
5. The Wiener-Askey Polynomial Chaos for Stochastic Differential Equations - DONGBIN XIU AND GEORGE EM KARNIADAKIS - 2002
6. 多项式混沌展开和最大熵原理的结构动力特性不确定性量化 - 万华平 - 2019
7. 基于稀疏多项式混沌展开的孤岛微电网全局灵敏度分析 - 王晗 - 2019
8. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence - Anneli Sch€oniger - 2014
9. Bayesian Model Selection and Model Averaging - Larry Wasserman - 2000
10. Maximum likelihood Bayesian averaging of uncertain model predictions - S. P. Neuman - 2003
11. Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models - RANGASAMI L. KASHYAP - 1982
12. The Bayesian information criterion: background, derivation, and applications - Andrew A. Neath - 2012
13. Multimodel Inference: Understanding AIC and BIC in Model Selection - Burnham, K.P. and Anderson, D.R. - 2004