А. В. Бернштейн (Москва, ИСА РАН)**. Аппроксимация зависимостей как задача снижения размерности.**

Рассмотрим задачу построения зависимости $y = f_N(x)$ для неизвестной функции $y = f(x)$, $x \in \mathbf{R}^p$, $y \in \mathbf{R}^q$, по заданному множеству ее известных значений $D_N = \{(x_i, y_i = f(xi)),\ i = 1, 2, \ldots, N\}$, минимизирующей среднеквадратическую ошибку аппроксимации

$$\varepsilon(f_N|D_N) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|f_N(x_i) - f(x_i)|^2}. \tag{1}$$

Функция $f(x)$ определяет в пространстве векторов $(x, y) \in \mathbf{R}^{p+q}$ p-мерное многообразие $M(f) = \{(x, y)\colon\ y = f(x),\ x \in \mathbf{R}^p\}$. Определяемое аналогичным образом при помощи аппроксимирующей зависимости $f_N(x)$ многообразие $M(f_N)$ аппроксимирует как многообразие $M(f)$, так и множество D_N: величина ε^2 (1) равна среднему квадрату расстояний между точками $(x_i, y_i = f_N(x_i)) \in M(f)$ множества D_N и точками $(x_i, y_i = f_N(x_i)) \in M(f_N)$.

Пусть $Z_N = \{z_i \in \mathbf{R}^n,\ i = 1, 2, \ldots, N\}$ — заданное множество n-мерных векторов, $\Sigma = \{m, C_m, R_m\}$ — процедура снижения размерности до размерности $m < n$, построенная по множеству Z_N и определяемая преобразованиями сжатия C_m: $z \in \mathbf{R}^n \ \rightarrow\ \lambda = C_m(z) \in \mathbf{R}^m$ и восстановления R_m $\lambda \in \mathbf{R}^m \ \rightarrow\ z = R_m(\lambda) \in \mathbf{R}^n$. Процедура Σ должна обеспечивать среднеквадратическую близость

$$\delta(\Sigma|Z_N) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|z_i - R_m(C_m(z_i))\|^2} \tag{2}$$

между исходными и восстановленными векторами. Процедура Σ определяет в $\mathbf{R}^n$ m-мерное параметрическое многообразие $M(Z_N) = \{R_m(\lambda) \in \mathbf{R}^n\colon\ \lambda \in \mathbf{R}^m\}$, аппроксимирующее множество Z_N: величина δ^2 (2) является усреднением квадратов расстояний между исходными точками $z_i \in Z_N$ и точками $R_m(\lambda_i)$, лежащими на многообразии $M(Z_N)$, соответствующими значениям $\lambda_i = C_m(z_i)$ параметра λ. Заметим, что при фиксированной процедуре восстановления R_m наилучшая процедура сжатия $C_m(z)$ состоит в проектировании вектора z на многообразие $M(Z_N)$: $C_m(z) = \arg\min_\lambda \|z - R_m(\lambda)\|$.

Рассмотрим D_N как множество $z_i = (x_i, y_i = f(xi)) \in \mathbf{R}^{p+q}$ и построим для него процедуру снижения размерности $\Sigma = \{m, C_m, R_m\}$, обеспечивающую заданную точность δ (2) и определяющую соответствующее m-мерное многообразие $M(D_N)$. В качестве нормы в (2) векторов $z = (x, y) \in \mathbf{R}^{p+q}$ выберем $\|(x, y)\|_A = (A^2\|x\|^2 + \|y\|^2)^{1/2}$. В силу исходной факторизации $z = (x, y)$, процедура восстановления определяет факторизацию $R_m(\lambda) = (x(\lambda); y(\lambda)) = (R_{mx}(\lambda) = x(\lambda); R_{my}(\lambda) = y(\lambda))$.

Будем рассматривать задачу аппроксимации как задачу построения по заданному вектору $x \in \mathbf{R}^p$ такого вектора $y = f_N(x) \in \mathbf{R}^q$, что $(p + q)$-мерный вектор $z(x) = (x; fN(x))$ является ближайшим к многообразию $M(D_N)$, аппроксимирующему множество данных D_N. Точное решение этой задачи дается формулой $y = f_N(x) = R_{my}(\arg\min_\lambda \|x - R_{mx}(\lambda)\|_A)$, где вес A оптимизируется для минимизации ошибки аппроксимации (1).

Заметим, что размерность m сжатого вектора может быть меньше размерности p вектора аргументов x, и p-мерное многообразие $M(f)$ будет аппроксимироваться многообразием $M(f_N)$ меньшей размерности. Тем самым, автоматически решается задача построения аппроксимирующей зависимости в ситуации, когда имеется зависимость среди компонентов x, или функция $f(x)$ зависит от x только через вектор-функцию меньшей размерности. В линейных задачах регрессионного, ковариационного и факторного анализа для близких целей фактически рассматриваются линейные процедуры снижения размерности.