А. В. Бернштейн (Москва, ИСА РАН). **Условие разделимости многообразия данных в задаче снижения размерности.**

Содержательно задача снижения размерности состоит в построении по множеству данных $X_{(N)} = \{X_i \in \mathbf{R}^p,\ i = 1, 2, \ldots, N\}$ процедуры $\Sigma = \{q, C_q, R_q\}$, определяемой размерностью q сжатых данных, преобразованием сжатия C_q: $X \in \mathbf{R}^p \to y = C_q(X) \in \mathbf{R}^q$ и преобразованием восстановления R_q: $y \in \mathbf{R}^q \to X = R_q(y) \in \mathbf{R}^p$. Если X — исходный вектор, а $X^* = R_q(C_q(X))$ — восстановленный вектор, получающийся путем последовательного применения процедур сжатия и восстановления, то процедура Σ должна обеспечивать близость между векторами X и X^*, и качество процедуры Σ может быть оценено по выборке средней ошибкой восстановления $\varepsilon(\Sigma) = (N^{-1} \sum_{i=1}^{N} \|X_i - X_i^*\|^2)^{1/2}$, вычисленной по множеству данных $X_{(N)}$.

Очевидно, что в такой постановке задача бессмысленна, и легко построить процедуру Σ^* с $q = 1$, обеспечивающую нулевую ошибку восстановления $\varepsilon(\Sigma^*)$ для множества данных $X_{(N)}$. Поэтому строгая математическая постановка задачи предполагает обеспечение близости $X \approx X^*$ не только для точек из $X_{(N)}$, но и для новых (out-of-sample) точек, что, в свою очередь, предполагает наличие модели данных, определяющей механизм порождения данных.

Общепринятая модель данных задается неизвестной гладкой вектор-функцией

$$X = F(b) = (F_1(b), F_2(b), \ldots, F_p(b))^T \in \mathbf{R}^p, \quad b \in \mathbf{B}, \tag{1}$$

отображающей открытое множество $\mathbf{B} \subset \mathbf{R}^q$ в $\mathbf{R}^p$, и при этом функция $F(b)$ не принимает одинаковых значений для различных значений аргумента $b \in \mathbf{B}$. Иногда в модель (1) добавляется случайный аддитивный шум. Предполагается, что множество $X_{(N)}$ есть множество значений функции (1) в точках $\{b_1, b_2, \ldots, b_N\} \subset \mathbf{B}$, которые выбираются в $\mathbf{B}$ случайно в соответствии с неизвестным распределением, носитель которого совпадает с $\mathbf{B}$.

Между отображениями C_q и R_q имеется тесная взаимосвязь. Если задано отображение восстановления R_q, определяющее q-мерное многообразие данных

$$\mathbf{X} = \mathbf{X}_F(\mathbf{B}) = \{X = F(b) \in \mathbf{R}^p\colon\ b \in \mathbf{B} \subset \mathbf{R}^q\} \tag{2}$$

в $\mathbf{R}^p$, то наилучшее преобразование сжатия C_q есть проекция на это многообразие: $C_q(X) = \arg\min_{y \in \mathbf{Y}} \{\|X - R_q(y)\|\}$, где $\mathbf{Y}$ — образ пространства $\mathbf{X}$ отображения C_q.

Если задано отображение сжатия C_q, определяющее образ $Y_{(N)} = \{y_1, y_2, \ldots, y_N\}$ множества $X_{(N)}$ при отображении C_q, то отображение R_q есть результат решения задачи восстановления неизвестного отображения $\mathbf{Y}$ в $\mathbf{R}^p$ по данным $\{(y_i = C_q(\mathrm{X}_i), X_i),\ i = 1, 2, \ldots, N\}$, состоящим из значений неизвестного отображения в точках множества $Y_{(N)}$. Очевидно, что для возможности такого восстановления необходимо, чтобы все точки множества $Y_{(N)}$ были различны, или, по крайней мере, одинаковые значения переменных y_i и y_j соответствовали близким значениям X_i и X_j.

Многие процедуры снижения размерности основаны на линейных преобразованиях сжатия $C_q = C_{q,h} = h^T \times X$, определяемых матрицами h размера $p \times q$, состоящими из q линейно независимых столбцов $h_1, h_2, \ldots, h_q \in \mathbf{R}^p$. Обозначим $X(h) = h \times h^T \times X$ проекцию вектора $X \in \mathbf{X}$ на линейное q-мерное подпространство $L(h)$, натянутое на столбцы матрицы h, и пусть $\mathbf{X}_F(h)$ — образ многообразия $\mathbf{X}$ при этом проектировании.

Многообразие данных $\mathbf{X}$ вида (2) удем называть *h-разделимым*, если отображение проектирования является взаимно однозначным, а матрицу h в этом случае будем называть *h-разделяющей*. Очевидно, что h-разделимость многообразия данных $\mathbf{X}$ обеспечивает возможность построения отображения восстановления R_q для заданного отображения сжатия $C_{q,h}$.

Теорема. *Для того чтобы матрица h обеспечивала h-разделимость многообразия данных $\mathbf{X}$, необходимо и достаточно, чтобы для всех $b \in \mathbf{B}$ матрицы*

$S(h, b) = h^T \times D_F(b)$ были невырожденными. Здесь $D_F(b)$ есть якобиан отображения F (1), строчками которого являются транспонированные векторы градиентов $\nabla^T F_j(b)$ компонент функции $F(b)$, $j = 1, 2, \ldots, p$.