L'apprentissage face à la malédiction de la grande dimension

Collège de France

Cours 4: **Réduction de dimensionnalité et débruitage**

Stéphane Mallat

1 Réduction de la dimensionnalité

Nous avons vu que le fait que x appartienne à un ensemble de grande dimension d implique que l'approximation de fonctions $f(x)$ localement régulières souffre de la malédiction de la dimensionnalité. Le nombre d'examples doit augmenter exponentiellement avec la dimension.

Pour contourner ce problème, on peut espérer que les données x appartiennent à un ensemble Ω dont la dimension intrinsèque est plus petite que la dimension d de l'espace ambiant. Autrement dit, que les données x aient elles-même une forme de régularité qui implique que le nombre de degrés de libertés de x est nettement inférieur à d. Un signal, que ce soit une image, un son ou toutes autres données, a en effet des formes de régularité qui interdisent à ses d variables de varier librement. L'enjeux ici va être de comprendre comment mettre en évidence cette réduction de dimensionalité, et l'utiliser pour les applications.

Nous allons maintenant considérer $x(u)$ comme une fonction de $u \in \mathbb{R}^\ell$ avec $1 \leq \ell \leq 4$. Pour un son $\ell = 1$ et u est le temps, pour une image $\ell = 2$ et $u = (u_1, u_2)$ est une variable d'espace. Pour une vidéo ou un bloque de données tridimensionnelles $\ell = 3$, et $\ell = 4$ si le bloque de données tridimensionelles varie dans le temps. Nous restons cependant en basse dimension car $\ell \leq 4$ et le plus souvent $\ell \leq 2$.

Au-delà de l'apprentissage, réduire la dimension d'un signal $x(u)$ est au coeur des problématiques du traitement du signal, pour la compression de signaux, pour le débruitage ou pour les problèmes inverses. Nous allons nous poser le même type de questions que celles posées précédemment, mais la fonction $f(x)$ avec $x \in \mathbb{R}^d$ en grande dimension est remplacée par $x(u)$ avec $u \in \mathbb{R}^l$ en basse dimension. On va essayer d'approximer $x(u)$ avec un nombre minimum de variables tout en introduisant une petite erreur. Il s'agit donc à nouveau d'un problème d'approximation, mais en basse dimension.

1.1 Approximation dans une base

La représentation de signaux dans des bases orthonormales permet d'exprimer simplement les problèmes de réduction de dimensionnalité, et offre déjà une grande flexibilité. On distinguera des approximations *linéaires* et *non linéaires*. Les outils de base du traitement du signal, comme les filtres sont souvent des opérateurs linéaires. Cependant, l'utilisation d'opérateurs non-linéaires, comme les opérateurs de seuillage dans des bases, peuvent parfois faire beaucoup mieux. On verra notamment apparaître la notion de parcimonie, qui est une forme duale de régularité dans un cadre non-linéaire.

On note le produit scalaire de deux éléments de $\mathbb{C}^d$

$$\langle x, y \rangle = \sum_{u=1}^{d} x(u) y(u)^*$$

et la norme Euclidienne $\|x\|^2 = \sum_{u=1}^{d} |x(u)|^2$. On considère une base orthonormale $\mathcal{B} = \{g_k\}_{1 \leq k \leq d}$ de $\mathbb{C}^d$. Cela signifie que $\forall 1 \leq k, k' \leq d$,

$$\langle g_k, g'_k \rangle = \begin{cases} 1 \text{ si } k = k' \\ 0 \text{ sinon} \end{cases}$$

N'importe quel $x \in \mathbb{C}^d$ peut alors se décomposer dans la base $\mathcal{B} = \{g_k\}_{1 \leq k \leq d}$ sous la forme

$$x = \sum_{k=1}^{d} \langle x, g_k \rangle \, g_k \quad \text{et} \quad \|x\|^2 = \sum_{k=1}^{d} |\langle x, g_k \rangle|^2.$$

1.2 Approximations linéaires

Une première approche *linéaire* pour réduire la dimensionalité consiste à ne garder que les M premiers coefficients de x dans la base. On note alors x_M cette approximation, qui est une approximation de x de dimension M :

$$x_M = \sum_{k=1}^{M} \langle x, g_k \rangle g_k.$$

L'approximation x_M est la projection orthogonale de x sur l'espace $V_M = Vect(\{g_k\}_{1 \leq k \leq M})$ générée par les M premiers vecteurs de la base. L'erreur est

$$\epsilon_M^2 = \|x - x_M\|^2 = \sum_{k=M+1}^{d} |\langle x, g_k \rangle|^2$$

Si la base $\mathcal{B}$ est la base de Fourier alors x_M est une représentation du signal x qui ne garde que ses M plus basses fréquences.

Pour que l'erreur ϵ_M soit petite il faut que les coefficients $(\langle x, g_k \rangle)_{k>M}$ éliminés soient petits. Sachant que $\|x\|^2 = \sum_k |\langle x, g_k \rangle|^2$ il faut que l'amplitude des coefficients $|\langle x, g_k \rangle|$ décroisse rapidement lorsque k augmente. La proposition suivante donne une majoration de l'erreur d'approximation dans le cas où les coefficients $|\langle x, g_k \rangle|$ décroissent comme $k^{-\alpha}$.

Proposition 1 *Si $|\langle x, g_k \rangle| \leq C k^{-\alpha}$ avec $\alpha > 1/2$ alors l'erreur ϵ_M vérifie*

$$\epsilon_M^2 \leq \frac{C^2}{2\alpha - 1} \, M^{1-2\alpha} \ .$$

Démonstration : On a

$$\begin{aligned} \epsilon_M^2 &= \sum_{k=M+1}^{d} |\langle x, g_k \rangle|^2 \leq C^2 \sum_{k=M+1}^{d} k^{-2\alpha} \\ &\leq C^2 \sum_{k=M+1}^{d} \int_{k-1}^{k} \frac{1}{x^{2\alpha}} \mathrm{d}x = C^2 \int_{M}^{d} \frac{1}{x^{2\alpha}} \mathrm{d}x \leq C^2 \int_{M}^{+\infty} \frac{1}{x^{2\alpha}} \mathrm{d}x \leq \frac{C^2}{2\alpha - 1} \, M^{1-2\alpha} \ . \end{aligned}$$

□

Si x est la réalisation d'un vecteur aléatoire X on peut essayer de trouver la base orthonormale $\mathcal{B}$ qui minimise l'erreur ϵ_M^2 en moyenne sur les realisations x de X. Nous verrons ultérieurement que cette base est une base orthonormale qui diagonalise la matrice de covariance de X, en rangeant les vecteurs g_k dans l'ordre décroissance des valeurs propres de la covariance. On appelle cette base une base de Karhunen-Loève, ou une base de composantes principales. Si l'on sait que le processus X est stationnaire (modulo d), on verra que cette base est la base de Fourier.

Nous verrons qu'approximer un signal x dans une base de Fourier est équivalent à approximer ce signal avec un échantillonnage uniforme. Ce type d'approximation est efficace si le signal est uniformément régulier. Ce n'est par contre pas le cas pour des signaux réguliers par morceaux, qui incluent des discontinuités. Ce type de signaux réguliers par morceaux donnent envie de définir un échantillonnage *adaptatif non linéaire*, où l'on échantillonnera avec plus de points sur les zones irrégulières et moins de points sur les zones régulières. Choisir de façon optimale la position de ces échantillonns est un problème difficile qui demande d'évaluer la régularité locale du signal. Une approche plus simple et efficace consiste à trouver une base dans laquelle l'amplitude des coefficients de x reflète la régularité locale de x. On peut alors approximer le signal en choissant les plus grands coefficients dans la base, ce qui correspond à une approximation non-linéaire.

1.3 Approximations non-linéaires

Nous allons approximer x par une combinaison linéaire de M vecteurs dans une base $\mathcal{B} = \{g_k\}_{1 \leq k \leq d}$

$$x_M = \sum_{k \in I_M} \langle x, g_k \rangle \, g_k \ ,$$

en nous laissant la possibilité de choisir librement l'ensemble I_M de ces M vecteurs en fonction de x. Le choix de ces vecteurs n'est donc pas fixé à priori, comme dans le cas linéaire.

Le meilleur choix d'indices I_M est celui qui minimise l'erreur d'approximation

$$\epsilon_M^2 = \|x - x_M\|^2 = \sum_{k \notin I_M} |\langle x, g_k \rangle|^2.$$

Pour cela il faut que I_M soit l'ensemble des indices k des M plus grands coefficients $|\langle x, g_k \rangle|$. Cela revient à ne garder que les coefficients plus grand qu'un seuil T_M ajusté pour qu'il n'y en ait que M :

$$I_M = \{k \, / \, |\langle x, g_k \rangle| \geq T_M\} \ .$$

Comme $\|x\|^2 = \sum_k |\langle x, g_k \rangle|^2$, pour que l'erreur ϵ_M soit petite, il faut qu'il y ait un petit nombre de grands coefficients $|\langle x, g_k \rangle|^2$ qui absorbent l'essentiel de "l'énergie" $\|x\|^2$ de x, tandis que les autres ont une amplitude négligeable. Cela signifie que le signal x a une représentation *parcimonieuse* dans la base $\mathcal{B}$, autrement dit qu'il y a un petit nombre de coefficients non-négligeables. Contrairement au cas linéaire, la position des grands coefficients peut varier arbitrairement en fonction de x. La parcimonie n'impose pas que la position des grands coefficients soit toujours la même. Ainsi, on verra que si $\mathcal{B}$ est une base orthonormale d'ondelettes, les grands coefficients sont prés des singularités et leur position va donc dépendre de la position de ces singularités.

2 Application au débruitage

2.1 Model de signal et de bruit

On considère que la donnée $x \in \mathbb{R}^d$ est contaminée par un bruit additif B que l'on modélise comme un bruit blanc :

$$Z(u) = \underbrace{x(u)}_{déterministe} + \underbrace{B(u)}_{aléatoire}$$

Nous modélisons donc le signal contaminé Z comme la somme d'un signal sous-jacent x *déterministe* et d'un bruit B suivant un modèle *aléatoire*. Ce bruit peut provenir de capteurs, d'erreurs de transmission etc... En pratique, on arrive à construire un modèle aléatoire du bruit car sa structure est souvent bien plus simple que celle du signal. Il est en revanche beaucoup plus compliqué de construire un modèle stochastique pour des signaux comme des images, des sons ou tout autre. C'est pour cela que l'on se contentera d'adopter une modélisation déterministe du signal, qui spécifie plus simplement que le signal appartient à un ensemble particulier.

On suppose que le bruit B est un *bruit blanc* de variance σ^2. Cela signifie que les variable aléatoire $B(u)$ sont de moyenne nulle $\mathbb{E}[B(u)] = 0$, qu'elles sont non corrélées entre elles et de même variance σ^2 :

$$\forall u, u' \ , \ \mathbb{E}[B(u)B(u')^*] = \sigma^2 \delta[u - u'].$$

Soit $\mathcal{B} = \{g_k\}_{1 \leq k \leq d}$ une base orthonormée. Chaque coefficient de décomposition de B est une variable aléatoire

$$\langle B, g_k \rangle = \sum_{u=1}^{d} B(u) \, g_k^*(u) \ .$$

La proposition suivante montre que ces coefficients définissent toujours un bruit blanc.

Proposition 2 *Si $\mathcal{B} = \{g_k\}_{1 \leq k \leq d}$ une base orthonormée et B est un bruit blanc de variance σ^2 alors les d variables aléatoires $\{\langle B, g_k \rangle\}_{1 \leq k \leq d}$ définissent un bruit blanc de variance σ^2.*

Démonstration : Ces coefficients sont de moyenne nulle car $\mathbb{E}[B(u)] = 0$ et donc

$$\mathbb{E}[\langle B, g_k \rangle] = \sum_{u=1}^{d} g_k^*(u) \mathbb{E}[B(u)] = 0.$$

On vérifie que les coefficients $(\langle B, g_k \rangle)_{1 \leq k \leq M}$ sont décorrélés par un calcul direct :

$$\begin{aligned}
\mathbb{E}\left[\langle B, g_k \rangle \langle B, g'_k \rangle^*\right] &= \mathbb{E}\left[\sum_{u=1}^{d} B(u) g_k(u)^* \sum_{u'=1}^{d} B(u')^* g'_k(u')\right] \\
&= \sum_{u=1}^{d} \sum_{u'=1}^{d} g_k(u)^* g'_k(u') \mathbb{E}[B(u) B(u')^*] \\
&= \sum_{u=1}^{d} \sum_{u'=1}^{d} g_k(u)^* g'_k(u') \sigma^2 \delta[u - u'] \\
&= \sigma^2 \sum_{u=1}^{d} g'_k(u) g_k(u)^* = \sigma^2 \langle g'_k, g_k \rangle = \sigma^2 \delta[k - k'].
\end{aligned}$$

□

On calcule un estimateur $\tilde{X}$ de x, en appliquant un opérateur L sur les données contaminées Z : $\tilde{X} = LZ$. L'objectif est de minimiser le risque, ici défini comme étant l'erreur quadratique moyenne :

$$R = \mathbb{E}\left[\|\tilde{X} - x\|^2\right]$$

2.2 Estimation linéaire et dilemme biais-variance

Nous allons d'abord considérer une estimation *linéaire* dans une orthonormal base $\mathcal{B} = \{g_k\}_{1 \leq k \leq d}$. Si on sait *à priori* que le signal x est bien approximé par sa projection dans l'espace $V_M = Vect\{g_k\}_{1 \leq k \leq M}$ généré par les M premiers vecteurs, alors on peut réduire le bruit en effectuant une projection orthogonale des données bruitées sur V_M :

$$\tilde{X} = P_{V_M} Z = \sum_{k=1}^{M} \langle Z, g_k \rangle \, g_k \ . \tag{1}$$

La proposition suivante montre que l'erreur de cette estimation, tout comme dans l'apprentissage supervisé, se décompose comme la somme d'un terme de biais, autrement dit d'erreur de modèle, plus un terme ici de variance qui mesure la fluctuation du bruit résiduel aprés projection. Celui-ci est proportionel à la dimension M de l'espace V_M.

Proposition 3 *Si V_M est un espace de dimension M alors*

$$R = \mathbb{E}\left[\|\tilde{X} - x\|^2\right] = \|x - P_{V_M} x\|^2 + M\sigma^2 \ .$$

Démonstration : Comme $P_{V_M} Z = P_{V_M} x + P_{V_M} B$,

$$\|x - \tilde{X}\|^2 = \|x - P_{V_M} Z\|^2 = \|x - P_{V_M} x - P_{V_M} B\|^2 = \|x - P_{V_M} x\|^2 + \|P_{V_M} B\|^2$$

car $x - P_{V_M} x \in V_M^{\perp}$ et $P_{V_M} B \in V_M$ et sont donc orthogonaux. D'où :

$$R = \mathbb{E}\left[\|\tilde{X} - x\|^2\right] = \|x - P_{V_M} x\|^2 + \mathbb{E}\left[\|P_{V_M} B\|^2\right]$$

Soit $\mathcal{B} = \{g_k\}_{1 \leq k \leq M}$ une base orthonormée de V_M,

$$\mathbb{E}\left[\|P_{V_M} B\|^2\right] = \mathbb{E}\left[\sum_{k=1}^{M} |\langle B, g_k \rangle|^2\right] = M \, \sigma^2.$$

où la dernière égalité se déduit de la Propositon 2, ce qui conclut la preuve. □

On se retrouve donc une nouvelle fois dans le cadre du dilemme biais variance. La variable M est ici la dimension du modèle linéaire. En apprentissage, cette variable est remplacé par le terme de complexité $\log(|\mathcal{H}|)$ de la classe d'approximation. Plus le nombre de paramètres M est grand, plus l'erreur de modèle sera faible (espace d'approximation V_M de plus en plus grand) mais plus grande sera la variance du bruit. L'erreur totale est minimum, à un facteur 2 prés, pour un M pour lequel les deux termes sont du même ordre.

Le dilemne bias-variance s'illustre par l'ajout d'un bruit blanc sur un signal régulier. Puisque le bruit est de moyenne nulle, on a envie de moyenner le signal pour atténuer le bruit. Il faut alors déterminer la taille de la fenêtre de moyennage. Si cette fenêtre est petite, elle moyenne le sinal sur peu de coefficients ce qui n'enlèvera pas beaucoup de bruit. Si cette fenêtre est grande, on moyenne beaucoup de valeurs et on réduit donc le bruit mais également certaines composantes de hautes fréquences du signal. Ajuster la taille de la fenêtre est un dilemne de bias-variance.

2.2.1 Débruitage par seuillage non linéaire

Lorsque le signal n'est pas uniformément régulier mais inclut des composantes transitoires ou des discontinuités, celles-ci vont être fortement dégradées par un moyennage. Dans le cas d'une image, cela va restaurer une image floue. Pour améliorer ce type d'estimation il faut détecter et distinguer les parties irrégulieres liées au signal sous-jacent et celles liées au bruit, sachant que l'on ne connait pas a priori la localisation des irrégularités du signal. Au lieu de calculer une estimation linéaire comme en (1) par projection sur les M premiers vecteurs d'une base $\mathcal{B}$, qui ne dépendent pas du signal Z, on va sélectionner les vecteurs de $\mathcal{B}$ en fonction de Z.

Chaque coefficient du bruit se décompose sous la forme

$$\langle Z, g_k \rangle = \langle x, g_k \rangle + \langle B, g_k \rangle.$$

Idéalement on voudrait ne garder que les coefficients pour lesquels $|\langle x, g_k \rangle| \geq |\langle B, g_k \rangle|$. Cependant on ne connait pas les valeurs de $|\langle x, g_k \rangle|$. On peut par contre estimer la valeur maximum que les coefficients du bruit peuvent atteindre

$$T_m \geq \max_{1 \leq k \leq d} |\langle B, g_k \rangle|.$$

Si $|\langle Z, g_k \rangle| > \alpha\, T_m$ alors nécessairement $|\langle x, g_k \rangle| > (1 - \alpha) T_m$. Les algorithmes de seuillage consistent à ne garder que les coefficients au-dessus d'un seuil $T = \alpha T_m$:

$$\tilde{X} = L\, Z = \sum_{k=1}^{d} \rho_T(\langle Z, g_k \rangle)\, g_k$$

où le seuillage est défini par

$$\rho_T(u) = \begin{cases} 0 & si |u| < T \\ u & si |u| \geq T \end{cases}.$$

Cet estimateur de seuillage est non-linéaire et s'adapte aux propriétés du signal x. Dans le cas où B est un bruit blanc Gaussien, on peut démontrer que $T_m \sim \sigma \sqrt{2 \log d}$ et les valeurs optimales du seuil T sont obtenues pour $\alpha \leq 1$ car trés peu de coefficients du bruit atteignent cette valeur extrême.

Lorsque la base $\mathcal{B}$ est une base d'ondelettes, nous verrons que ces estimateurs de seuillages effectuent une régularisation adaptative qui moyenne le signal bruité Z sur des grands domaines lorsque le signal sous-jacent x est régulier, et qui limite ce moyennage dans les zones où x est irrégulier.