Introduction à l'apprentissage supervisé

Y. Pradat (*auteur*), L. Verlingue & D. Gautheret (*responsables cours*)

CentraleSupelec (labo MICS) & Institut Gustave Roussy (IGR)

23 janvier 2020

Notations

Notations

- $n, p \in \mathbb{N}^*$ nombre d'observations (unités, individus) et nombre de variables (facteurs, cofacteurs) ;
- X (resp $\mathbf{X}$) variable aléatoire scalaire (resp vectorielle) ;
- $\mathbf{X}_{\mathbf{1:n}}$ vecteur (resp matrice) des variables aléatoires $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ (resp $\mathbf{X}_\mathbf{1}, \ldots, \mathbf{X}_\mathbf{n}$) ;
- x (resp $\boldsymbol{x}$) observation de X (resp $\mathbf{X}$) ;
- $\boldsymbol{x}_{\mathbf{1}:\boldsymbol{n}}$ observation de $\mathbf{X}_{\mathbf{1:n}}$.

Objectif :

1. Proposer un modèle mathématique ;
2. Estimer les paramètres du modèle.

Definition vraisemblance

Définition 1. Modèle statistique

Un modèle statistique est une collection de lois (ou densités) candidates paramétrée par $\theta \in \Theta$

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\} \tag{1}$$

Définition 2. Vraisemblance

Soient $\mathbf{x_{1:n}}$ *un échantillon d'observations de* $\mathbf{X_{1:n}} \sim p_{\theta^*}$. *La vraisemblance du paramètre* θ *pour cet échantillon est*

$$\mathcal{L}(\theta; \mathbf{x}_{1:n}) = \prod_{i=1}^{n} p_\theta(x_i) \tag{2}$$

Definition vraisemblance

Définition 1. Modèle statistique

Un modèle statistique est une collection de lois (ou densités) candidates paramétrée par $\theta \in \Theta$

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\} \tag{1}$$

Définition 2. Vraisemblance

Soient $\boldsymbol{x}_{\mathbf{1:n}}$ *un échantillon d'observations de* $\mathbf{X}_{\mathbf{1:n}} \sim p_{\theta^*}$. *La vraisemblance du paramètre* θ *pour cet échantillon est*

$$\mathcal{L}(\theta; \mathbf{x}_{1:n}) = \prod_{i=1}^{n} p_\theta(x_i) \tag{2}$$

Exemple 1

Soit $\boldsymbol{x}_{\mathbf{1:n}}$ *un n-échantillon de* $\mathbf{X}_{\mathbf{1:n}} \sim \mathcal{N}(\mu^*, {\sigma^2}^*)$. *Alors, la vraisemblance de* (μ, σ^2) *est*

$$\mathcal{L}(\mu, \sigma^2; \boldsymbol{x}_{\mathbf{1:n}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \tag{3}$$

Estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x}_{\mathbf{1}:\boldsymbol{n}}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x}_{\mathbf{1}:\boldsymbol{n}}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{L}(\theta; \boldsymbol{x}_{\mathbf{1}:\boldsymbol{n}}) \tag{4}$$

définit un estimateur du maximum de vraisemblance de θ^**.*

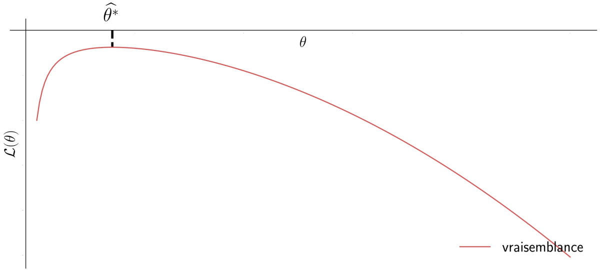
Estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x_{1:n}}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x_{1:n}}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta; \boldsymbol{x_{1:n}}) \tag{4}$$

définit un estimateur du maximum de vraisemblance de θ^**.*



Exemple estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x}_{\mathbf{1:n}}$ de $\mathbf{X}_{1:n} \sim p_{\theta^*}$, l'équation

$$\widehat{\theta^*}(\boldsymbol{x}_{\mathbf{1:n}}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta; \boldsymbol{x}_{\mathbf{1:n}}) \tag{5}$$

définit un estimateur du maximum de vraisemblance de θ^*.

Exemple 2

Soit $\boldsymbol{x}_{\mathbf{1:100}}$ un 100-échantillon de $\mathbf{X}_{1:100} \sim \mathcal{B}(\theta^*)$ (100 lancés de pièces identiques). L'estimateur par MV de θ maximise

$$\prod_{i=1}^{100} p_{\mathrm{X_i}}(x_i; \theta) \tag{6}$$

Exemple estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x}_{1:n}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x}_{1:n}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta; \boldsymbol{x}_{1:n}) \tag{5}$$

définit un estimateur du maximum de vraisemblance de θ^**.*

Exemple 2

Soit $\boldsymbol{x}_{1:100}$ *un* 100*-échantillon de* $\mathbf{X}_{1:100} \sim \mathcal{B}(\theta^*)$ *(100 lancés de pièces identiques). L'estimateur par MV de* θ *maximise*

$$\prod_{i=1}^{100} p_{\mathrm{X_i}}(x_i; \theta) \tag{6}$$

Pour une loi de Bernoulli, $p(x; \theta) = \theta^x (1-\theta)^{1-x}$ *de sorte que*

$$\hat{\theta^*}(\boldsymbol{x}_{1:100}) = \frac{1}{100} \sum_{i=1}^{100} x_i \tag{7}$$

Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

$$\forall i = 1, \ldots, n, \qquad \mathrm{Y}_\mathrm{i} = \boldsymbol{x_i}^\top \beta + \mathcal{E}_i \tag{8}$$

avec $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$ (résidus).

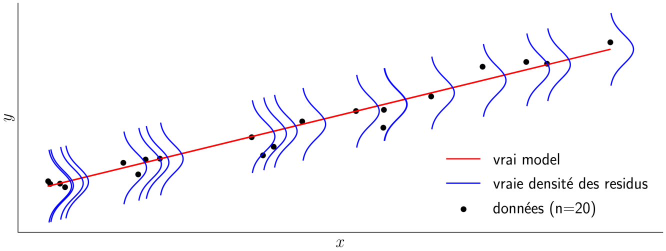
Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

$$\forall i = 1, \ldots, n, \qquad \mathrm{Y_i} = \boldsymbol{x_i}^\top \beta + \mathcal{E}_i \tag{8}$$

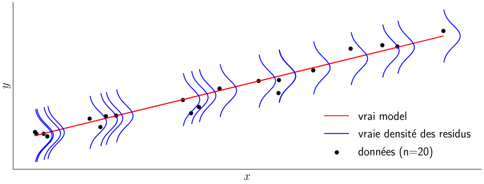
avec $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$ (résidus).

Explication



Construction modèle

Explication



- Les paramètres du modèle sont (β, σ^2).
- C'est un modèle statistique $\mathcal{M}_{\beta,\sigma^2}$ sur les densités conditionnelles $p_{\mathrm{Y_i}|\mathbf{x_i}=x_i}$.
- **Hypothèses du modèle**
  1. Relation linéaire entre $\mathbf{x_i}$ et $\mathrm{Y_i}$;
  2. Résidus indépendants et gaussiens ;
  3. Homoscedasticité.

Estimation paramètre β

Par maximum de vraisemblance,

$$\mathcal{L}(\beta, \sigma^2; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\boldsymbol{x_i}}(y_i) \tag{9}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \beta)^2} \tag{10}$$

Exercice : Calculer l'estimateur $\hat{\beta}(\boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}})$ de β par MV en minimisant $\ell(\beta, \sigma^2) = -\log \mathcal{L}(\beta, \sigma^2)$.

Estimation paramètre β

Par maximum de vraisemblance,

$$\mathcal{L}(\beta, \sigma^2; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\boldsymbol{x_i}}(y_i) \tag{9}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \beta)^2} \tag{10}$$

<u>Exercice</u> : Calculer l'estimateur $\hat{\beta}(\boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}})$ de β par MV en minimisant $\ell(\beta, \sigma^2) = -\log \mathcal{L}(\beta, \sigma^2)$.

Tous calculs faits, si $\boldsymbol{x_{1:n}^\top x_{1:n}}$ inversible,

$$\hat{\beta}(\boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \left(\boldsymbol{x_{1:n}^\top x_{1:n}}\right)^{-1} \boldsymbol{x_{1:n}^\top y_{1:n}} \tag{11}$$

Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

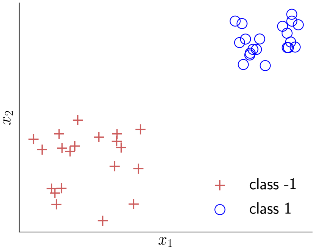
$$\forall i = 1, \ldots, n, \qquad \mathrm{Y}_\mathrm{i} \sim \mathcal{B}(\sigma(\boldsymbol{x_i}^\top \theta)) \tag{12}$$

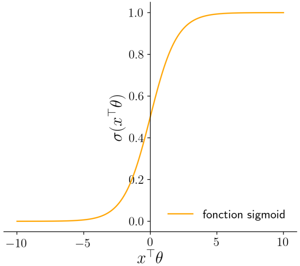
Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

$$\forall i = 1, \ldots, n, \qquad \mathrm{Y}_\mathrm{i} \sim \mathcal{B}(\sigma(\boldsymbol{x_i}^\top \theta)) \tag{12}$$

Explication





Estimation paramètre θ

Explications

- Les paramètres du modèle sont θ.
- C'est un modèle statistique $\mathcal{M}_\theta$ sur les densités conditionnelles $p_{\mathrm{Y_i}|\mathbf{X_i}=\boldsymbol{x_i}}$.
- **Hypothèses du modèle**
  1. Relation linéaire entre $\mathrm{logit}\, p_{\mathrm{Y_i}|\mathbf{X_i}=\boldsymbol{x_i}}(1)$ et $\boldsymbol{x_i}$;

Estimation Par maximum de vraisemblance,

$$\mathcal{L}(\theta; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\boldsymbol{x_i}}(y_i) \tag{13}$$

$$= \prod_{i=1}^{n} \sigma(\theta^\top \boldsymbol{x_i})^{y_i}(1-\sigma(\theta^\top \boldsymbol{x_i}))^{1-y_i} \tag{14}$$

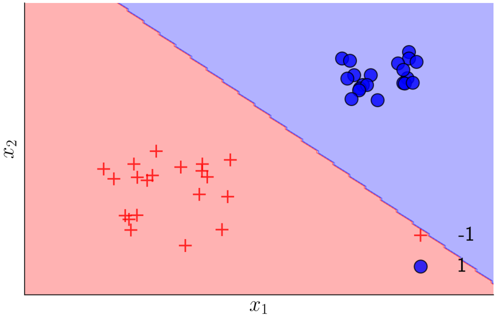
$$\ell(\theta; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top \boldsymbol{x_i}) + (1-y_i)\log(1-\sigma(\theta^\top \boldsymbol{x_i})) \tag{15}$$

Estimation paramètre θ

Comme $\sigma'(z) = \sigma(z)(1 - \sigma(z)$ on obtient que :

$$\nabla \ell(\theta) = -\sum_{i=1}^{n} \boldsymbol{x_i}(y_i - \sigma(\boldsymbol{x_i}^\top \theta)) \tag{16}$$

qui nous sert à minimiser $\ell(\theta)$ *numériquement*.



Modélisation de survie

Notations

1. $\mathrm{T_D}, \mathrm{T_C}$ variable aléatoires positives, temps à l'évènement et temps à la censure respectivement ;
2. Δ variable aléatoire binaire indiquant l'occurrence de l'évènement ;
3. $\mathrm{T} = \min(\mathrm{T_D}, \mathrm{T_C})$ variable aléatoire observée ;
4. $\mathbf{Z}$ vecteur aléatoire des covariables ;

Définition 4. Fonction de survie

La fonction de survie est donnée par

$$S : \begin{array}{ccc} \mathbb{R}_+ & \to & [0,1] \\ t & \mapsto & \mathbb{P}(\mathrm{T_D} \geq t), \end{array}$$

Modélisation de survie

Définition 5. Taux de risque

La taux de risque (instantané) est la fonction donnée par

$$\lambda \colon \begin{array}{ccc} \mathbb{R}_+ & \to & \mathbb{R}_+ \\ t & \mapsto & \lim_{h\to 0} \mathbb{P}(\mathrm{T_D} \leq t + h | \mathrm{T_D} \geq t). \end{array}$$

Modélisation de survie

Définition 5. Taux de risque

La taux de risque (instantané) est la fonction donnée par

$$\lambda\colon \begin{array}{ccc} \mathbb{R}_+ & \to & \mathbb{R}_+ \\ t & \mapsto & \lim_{h\to 0} \mathbb{P}(\mathrm{T_D} \leq t + h | \mathrm{T_D} \geq t). \end{array}$$

<u>Soit</u> : f_D (resp F_D) densité (resp f.r.) de $\mathrm{T_D}$. Alors

$$\lambda(t) = \frac{f_D(t)}{1 - F_D(t)} \tag{17}$$

$$= \frac{-S'(t)}{S(t)} \tag{18}$$

$$\text{d'où} \quad \int_0^t \lambda(s) ds = -\log(S(t)) \quad (\text{car } S(0)=1) \tag{19}$$

Construction modèle

Le modèle Soient $\delta_{1:n}, \boldsymbol{z_{1:n}}$ n-échantillon de $\Delta_{1:n}, \mathbf{Z_{1:n}}$. Le modèle de Cox modélise $\mathrm{T_{D,1:n}}$, à $\boldsymbol{z_{1:n}}$ fixés, via le taux de risque selon la relation

$$\lambda_i(t) = \lambda_0(t) e^{\boldsymbol{z_i}^\top \beta} \tag{20}$$

Construction modèle

Le modèle Soient $\delta_{1:n}, \boldsymbol{z_{1:n}}$ n-échantillon de $\Delta_{1:n}, \mathbf{Z_{1:n}}$. Le modèle de Cox modélise $\mathrm{T_{D,1:n}}$, à $\boldsymbol{z_{1:n}}$ fixés, via le taux de risque selon la relation

$$\lambda_i(t) = \lambda_0(t)e^{\boldsymbol{z_i}^\top \beta} \tag{20}$$

Explications **Hypothèses**

- *Proportionnalité* Exercice : Pourquoi “risques proportionnels” ?

Construction modèle

Le modèle Soient $\delta_{1:n}, \boldsymbol{z_{1:n}}$ n-échantillon de $\Delta_{1:n}, \mathbf{Z_{1:n}}$. Le modèle de Cox modélise $\mathrm{T_{D,1:n}}$, à $\boldsymbol{z_{1:n}}$ fixés, via le taux de risque selon la relation

$$\lambda_i(t) = \lambda_0(t) e^{\boldsymbol{z_i}^\top \beta} \tag{20}$$

Explications **Hypothèses**

- *Proportionnalité* Exercice : Pourquoi “risques proportionnels” ? Supposons $\boldsymbol{z} = 1$ pour le groupe traité et $\boldsymbol{z} = 0$ pour le groupe de contrôle. Alors,

$$\forall t \geq 0, \qquad \frac{\lambda(t,1)}{\lambda(t,0)} = \frac{\lambda_0(t) e^{\beta}}{\lambda_0(t)} = e^{\beta} \tag{21}$$

Construction modèle

Le modèle Soient $\delta_{1:n}, \boldsymbol{z_{1:n}}$ n-échantillon de $\Delta_{1:n}, \mathbf{Z_{1:n}}$. Le modèle de Cox modélise $\mathrm{T_{D,1:n}}$, à $\boldsymbol{z_{1:n}}$ fixés, via le taux de risque selon la relation

$$\lambda_i(t) = \lambda_0(t)e^{\boldsymbol{z_i}^\top \beta} \tag{20}$$

Explications **Hypothèses**

- *Proportionnalité* Exercice : Pourquoi "risques proportionnels" ? Supposons $\boldsymbol{z} = 1$ pour le groupe traité et $\boldsymbol{z} = 0$ pour le groupe de contrôle. Alors,

$$\forall t \geq 0, \qquad \frac{\lambda(t,1)}{\lambda(t,0)} = \frac{\lambda_0(t)e^{\beta}}{\lambda_0(t)} = e^{\beta} \tag{21}$$

- *Linéarité* entre $\log(\lambda_i)$ et les covariables $\boldsymbol{z_i}$.

Vraisemblance totale

Par maximum de vraisemblance

- $\mathbb{P}(\mathrm{T_i} = t_i | \Delta_i = 0) = \mathbb{P}(\mathrm{T_{D,i}} \geq t_i) = \mathcal{S}_i(t_i)$
- $\mathbb{P}(\mathrm{T_i} = t_i | \Delta_i = 1) = \mathbb{P}(\mathrm{T_{D,i}} = t_i) = \lambda_i(t_i)\mathcal{S}_i(t_i)$

Vraisemblance totale

Par maximum de vraisemblance

- $\mathbb{P}(\mathrm{T_i} = t_i | \Delta_i = 0) = \mathbb{P}(\mathrm{T_{D,i}} \geq t_i) = \mathcal{S}_i(t_i)$
- $\mathbb{P}(\mathrm{T_i} = t_i | \Delta_i = 1) = \mathbb{P}(\mathrm{T_{D,i}} = t_i) = \lambda_i(t_i)\mathcal{S}_i(t_i)$

Alors,

$$\begin{aligned}
\mathcal{L}(\beta; \mathbf{z_{1:n}}, \mathbf{t_{1:n}}, \delta_{\mathbf{1:n}}) &= \prod_{i=1}^{n} \lambda(t_i)^{\delta_i} S_i(t_i) \\
&= \prod_{i=1}^{n} \left[\frac{\lambda_i(t_i)}{\sum_{j \in R(t_i)} \lambda_j(t_j)} \right]^{\delta_i} [\sum_{j \in R(t_i)} \lambda_j(t_i)]^{\delta_i} S_i(t_i) \\
&= \prod_{i=1}^{n} \left[\frac{\lambda_0(t_i) e^{z_i^\top \beta}}{\sum_{j \in R(t_i)} \lambda_0(t_i) e^{z_j^\top \beta}} \right]^{\delta_i} [\sum_{j \in R(t_i)} \lambda_j(t_j)]^{\delta_i} S_i(t_i)
\end{aligned}$$

Vraisemblance partielle

On estimate β par

$$\widehat{\beta}(\mathbf{z_{1:n}}, \mathbf{t_{1:n}}, \delta_{\mathbf{1:n}}) \in \operatorname*{argmax}_{\beta} \mathcal{L}_{\text{partiel}}(\beta) = \prod_{i=1}^{n} \left[\frac{\lambda_0(t_i) e^{z_i^\top \beta}}{\sum_{j \in R(t_i)} \lambda_0(t_i) e^{z_j^\top \beta}} \right]^{\delta_i} \tag{22}$$

Exemple

individual	X_i	δ_i	Z_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

Exemple

individual	X_i	δ_i	Z_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

j	ordered failure time X_i	$\mathcal{R}(X_i)$	i_j	Likelihood contribution $\left[e^{\beta Z_i} / \sum_{j \in \mathcal{R}(X_i)} e^{\beta Z_j}\right]^{\delta_i}$
1	6	{1,2,3,4}	3	$e^{7\beta}/[e^{4\beta}+e^{5\beta}+e^{7\beta}+e^{3\beta}]$
2	8	{1,2,4}	2	1
3	9	{1,4}	1	$e^{4\beta}/[e^{4\beta}+e^{3\beta}]$
4	10	{4}	4	$e^{3\beta}/e^{3\beta}=1$

Nous avons vu 3 modèles différents :

1. Régression linéaire, paramètres β, σ^2. Formule théorique pour β (et σ^2).

Nous avons vu 3 modèles différents :

1. Régression linéaire, paramètres β, σ^2. Formule théorique pour β (et σ^2).
2. Régression logistique, paramètres θ. Estimation *numérique* (descente de gradient).

Nous avons vu 3 modèles différents :

1. Régression linéaire, paramètres β, σ^2. Formule théorique pour β (et σ^2).
2. Régression logistique, paramètres θ. Estimation *numérique* (descente de gradient).
3. Modèle de Cox, paramètres $\beta, \lambda_0(t)$. Estimation *numérique* de β (descente de gradient).