Probabilistic Graphical Models and Deep Generative Models

D'après Pierre Latouche et Pierre-Alexandre Mattei

16 octobre 2025

Table des matières

Résumé

mailto:pierre.latouche@uca.fr Page du cours sur https://lmbp.uca.fr/~latouche/mva/IntroductiontoProbabilisticGraphi
html

1 Notion de Modèle Graphique Orienté

1.1 Modèle Statistique

Définition 1.1 Un *échantillon* aléatoire $(X_1, \ldots, X_n)$ de taille n est un vecteur de n variables aléatoires indépendantes et identiquement distribuées suivant la même loi qu'une variable aléatoire X_0.

Définition 1.2 Un *modèle statistique* est une famille de lois de probabilités sur un espace χ. Le modèle est dénoté par :

$$\mathcal{M} = \{p(\cdot \mid \theta), \theta \in \Theta\}$$

- Si $\Theta \subseteq \mathbb{R}^d$, le modèle est dit *paramétrique*.
- Sinon, on dira que le modèle est *non paramétrique*.

Pour une variable aléatoire discrète X, on notera $p(x \mid \theta) = \mathbb{P}_\theta(X = x)$. Pour une variable aléatoire continue X, on notera $p(x \mid \theta) = f_\theta(x)$ où f_θ est une densité de probabilité. Dans la suite, on dira que X est aléatoire et que x est une réalisation ou observation de X (même chose pour X_i et x_i). Le modèle $\{p(x \mid \theta)\}$ est l'unique lien entre θ et l'observation x.

Définition 1.3 La *vraisemblance* (likelihood) de θ de l'observation x est la fonction :

$$l_x(\theta) = p(x \mid \theta)$$

La *log-vraisemblance* est la fonction :

$$L_x(\theta) = \log l_x(\theta)$$

Dans la suite, on va chercher à maximiser la vraisemblance par rapport au paramètre θ étant donné un échantillon et un modèle statistique. Puisqu'on a une hypothèse d'indépendance :

$$l_{(x_1,\dots,x_n)}(\theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

et on a :

$$L_{(x_1,\dots,x_n)}(\theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta)$$

La log-vraisemblance d'un échantillon est la somme des log-vraisemblances des variables partielles d'observation. C'est une fonction aléatoire (définissant une variable aléatoire) en θ.

Il y a trois versions principales du paramètre θ qui vont être utilisées :

1. θ^*, une constante inconnue qui est le véritable paramètre du modèle ;
2. θ_0, une constante connue qui est un candidat pour le paramètre θ et sert à tester une hypothèse sur la valeur de θ ;
3. $\hat{\theta}$, une variable aléatoire fonction de l'échantillon aléatoire, et qui sert à estimer θ.

1.2 Statistiques et Information

Définition 1.4 Une *statistique* T est une fonction de l'échantillon aléatoire $(X_1, \dots, X_n)$.

Par la suite, on supposera que les lois $p(x \mid \theta)$ dans le modèle ont même support et sont deux fois dérivables en θ, de sorte que les opérateurs de dérivation par rapport à θ et d'intégration par rapport à x commutent.

Définition 1.5 Le *score* d'un échantillon est défini par :

$$L'_{(X_1,\dots,X_n)}(\theta) = \sum_{i=1}^{n} L'_{X_i}(\theta)$$

C'est une fonction aléatoire de θ.

Théorème 1.1 En θ^*, l'espérance du score est nulle :

$$\mathbb{E}^*[L'_{(X_1,\dots,X_n)}(\theta^*)] = 0$$

Ici, $\mathbb{E}^*$ désigne l'espérance par rapport à la loi $p(\cdot \mid \theta^*)$, l'espérance par rapport à la vraie distribution. Dans la suite, sauf précisé autrement, c'est toujours par rapport à cette distribution qu'on calculera l'espérance.

Démonstration. Par définition d'une mesure de probabilité, on a :

$$\int l_x(\theta)\,\mathrm{d}x = \int p(x \mid \theta)\,\mathrm{d}x = 1$$

En dérivant sous le signe intégral :

$$\int l'_x(\theta)\,\mathrm{d}x = \int l_x(\theta)L'_x(\theta)\,\mathrm{d}x = 0$$

Par définition de $l_x(\theta)$ et par définition de l'espérance :

$$\mathbb{E}^*[L'_x(\theta^*)] = \int L'_x(\theta^*)p(x \mid \theta^*)\,\mathrm{d}x = 0$$

■

Théorème 1.2 L'information de Fisher de X en θ^* est :

$$I_X(\theta^*) = \mathbb{E}^*[-L''_X(\theta^*)] = \mathbb{V}^*(L'_X(\theta^*))$$

De même que précédemment, $\mathbb{V}^*$ désigne la variance calculée par rapport à la distribution θ^*.

Démonstration. Par définition :

$$L''_X(\theta) = \frac{l''_X(\theta)l_X(\theta) - l'_X(\theta)}{l_X(\theta)^2} = \frac{l''_X(\theta)}{l_X(\theta)} - \left(\frac{l'_X(\theta)}{l_X(\theta)}\right)^2$$

Donc, en $\theta = \theta^*$, et en passant à l'espérance :

$$I_X(\theta^*) = -0 + \mathbb{E}\left[\left(\frac{l'_X(\theta^*)}{l_X(\theta^*)}\right)^2\right] = \mathbb{E}\left[L'_X(\theta^*)^2\right] = \mathbb{V}\left(L'_X(\theta^*)\right)$$

■

Proposition 1.1 L'information de Fisher en θ^* pour un échantillon $(X_1, \dots, X_n)$ est $I_n(\theta)n \times I_{X_i}(\theta), \forall i$.

On notera $I_{X_i}(\theta) = I(\theta)$.

R On a $\mathbb{E}^*[L''_X(\theta^*)] < 0$, en supposant la fonction deux fois continuement dérivable sur un voisinage de θ^*, elle y est concave et donc θ^* est le lieu d'un maximum.

1.3 Estimation Fréquentistes

Définition 1.6 Un *estimateur* $\hat{\theta}(X_1, \dots, X_n)$ est une statistique. On définit le *biais* d'un estimateur par :

$$b_{\theta^*}(\hat{\theta}) = \mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] - \theta^*$$

Un estimateur est *non biaisé* si son biais est nul.

Définition 1.7 Un estimateur $\hat{\theta}$ est dit *asymptotiquement non-biaisé* si pour tout θ^* :

$$\mathbb{E}^*[\hat{\theta}(X_1, \dots, X_n)] \xrightarrow[n\to\infty]{} \theta^*$$

Définition 1.8 La *déviation carrée moyenne* d'un estimateur $\hat{\theta}$ est :

$$Q = \mathbb{E}^*[\left(\hat{\theta}\left(X_1, \dots, X_n\right) - \theta^*\right)^2]$$

Proposition 1.2 On a :

$$Q = b_{\theta^*}\left(\hat{\theta}\right)^2 + \mathbb{V}^*\left(\hat{\theta}\right)$$

Démonstration. C'est la formule de König-Huygens pour des variables i.i.d. ■

Dans la suite, on va s'intéresser à des estimateurs dont le comportement est asymptotiquement bon quand $n \to \infty$. On notera $\hat{\theta}\left(X_1, \dots, X_n\right) = \hat{\theta}_n$

Définition 1.9 Une suite d'estimateurs $\hat{\theta}_n$ est dite *consistente* quand $\hat{\theta}_n \xrightarrow[n\to\infty]{C} \theta^*$, pour une propriété de convergence C à préciser.

- On parle de *consistence forte* quand C est la convergence presque sûre ;
- On parle de *consistence en moyenne carrée* quand $Q(\hat{\theta}_n) \xrightarrow[n\to\infty]{} 0$;
- On parle de *consistence en probabilité* quand :

$$\forall \varepsilon > 0, \mathbb{P}\left(\theta^* - \varepsilon < \hat{\theta}_n < \theta^* + \varepsilon\right) \xrightarrow[n\to\infty]{} 1$$

En pratique, on utilise majoritairement la consistence en moyenne carrée, puisque faire tendre Q vers 0 revient à trouver un estimateur non-biaisé dont la variance tend vers 0, asymptotiquement.

Définition 1.10 L'*estimateur de vraisemblance maximale* (EVM, ou MLE (*maximum likelihood estimator*)) pour le modèle statistique $\mathcal{M} = \{p\left(\cdot \mid \theta\right) \mid \theta \in \Theta\}$ est :

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L_{(X_1,\dots,X_n)}\left(\theta\right) \qquad \text{(EVM)}$$

R Si $\hat{\theta}$ est un EVM de θ^* pour le modèle statistique $\mathcal{M}$, et ψ est une bijection, alors $\psi(\hat{\theta})$ est un EVM de $\psi(\theta^*)$.
Rien ne présuppose en général qu'un EVM existe ou est unique. On va donc souvent avoir besoin d'optimiser numériquement, sans garantie (en général) de trouver un maximum global. Par concavité de la vraisemblance dans un petit voisinage de θ^*, on sait que l'EVM y est unique.

1.4 Approche Bayésienne

Dans l'approche bayésienne, au lieu d'introduire des estimateurs, on introduit un loi $p(\theta)$ avant d'observer les données, puis on calcule la loi postérieure $p(\theta \mid x_1, \dots, x_n)$. Par la formule de Bayes :

$$\overbrace{p(\theta \mid x_1, \dots, x_n)}^{\text{Postérieur}} = \frac{\overbrace{p(x_1, \dots, x_n \mid \theta)}^{\text{Vraisemblance}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(x_1, \dots, x_n)}_{\text{marginalisation}}} \tag{1}$$

Cette formule s'applique notamment dans le cas de l'apprentissage supervisé.

1.4.0.1 Bernoulli et Beta

Faisons un petit exemple. On prend $\text{support}(X_i) = \{0, 1\}$, $p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$. On considère *a priori* $p(\mu) = Beta(\mu, a_0, b_0) = \frac{1}{B(a,b)} \mu^{a-1}(1-\mu)^{b-1} \forall \mu \in [0,1]$. On a $\mathbb{E}[\mu] = \frac{a}{a+b}$, $\mathbb{V}(\mu) = \frac{ab}{(a+b)^2(a+b+1)}$ et $\text{mode}(\mu) = \frac{a-1}{a+b-2}$.

Proposition 1.3 La distribution posétrieure de μ dans le modèle de Bernoulli est donnée par :

$$p(\mu \mid x_1, \dots, x_n) = Beta(\mu, a_n, b_n)$$

avec $a_n = a_0 + \sum_{i=1}^n x_i$ et $b_n = b_0 + n - \sum_i^n x_i$.

Définition 1.11 La distribution Beta est appelée la *distribution conjuguée a priori* de la loi de Bernoulli puisque la distribution postérieure est aussi une distribution Beta.

Dans le modèle Bayésien, l'estimateur a posteriori maximum est défini par :

$$\hat{\mu}_{\text{MAP}} = \frac{a_0 + \sum_{i=1}^n x_i - 1}{a_0 + b_0 + n - 2}$$

L'estimateur basé sur la distribution prédictive est :

$$\hat{\mu} = \frac{a_0 + \sum_{i=1}^n x_i}{a_0 + b_0 + n}$$

Proposition 1.4 On a : $\hat{\mu}_{\text{MAP}} \to \hat{\mu}_{\text{ML}}$ et $\hat{\mu} \to \hat{\mu}_{\text{ML}}$ quand $n \to \infty$.

Démonstration. ■

1.5 Modèles Graphiques

On cherche à donner des propriétés probabilistes sur les données, et particulièrement les modéliser par des distributions conditionnelles. Les modèles graphiques sont un formalisme permettant l'analyse de distributions conditionnelles grâce à de la théorie des graphes. On ne fera pas de rappels des notions de théorie des graphes dans ce polycopié, sauf pour introduire des notations. En général on considèrera des graphes orientés acycliques (GAO, ou DAG en anglais).

Définition 1.12 Un *modèle graphique* est une famille de distribution de probabilités factorisables d'une manière définie par un GAO donné. Chaque sommet du graphe va correspondre à une variable aléatoire, les arêtes correspondent à des dépendances entre variables.

Les modèles graphiques orientés correspondent aux modèles statistiques les plus communément utilisés.

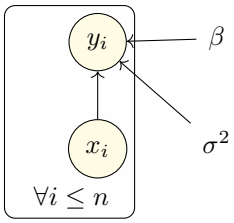
On va s'intéresser à la notion de factorisation dans un DAO :

Définition 1.13 Une distribution de probabilité p *se factorise dans* G lorsque pour tout x :

$$p(x) \prod_{i=1}^{d} p\left(x_i \mid x_{pa_i}\right)$$

où les pa_i parcourent l'ensemble des parents du sommet i.

Par exemple, le graphe suivant exprime la régression linéaire fréquentiste :



avec :

$$p\left(y_i \mid x_i, \beta, \sigma^2\right) = \mathcal{N}\left(y_i; {}^{\mathsf{t}}x_i\beta, \sigma^2\right)$$

2 Modèles de Mixtures Gaussiennes

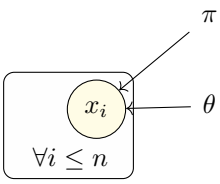
2.1 Mixtures Gaussiennes

Définition 2.1 Une *mixture gaussienne* à K composante est définie par la densité :

$$p\left(x \mid \pi, \theta\right) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x; \mu_k, \Sigma_k\right),$$

où $\theta = \left(\mu_k, \Sigma_k\right)_k$ et $\pi = {}^{\mathsf{t}}\left(\pi_1, \ldots, \pi_K\right) \in \Delta_K$ est le vecteur des poids.

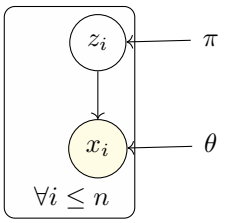
En modèle graphique :



Proposition 2.1 La log-vraisemblance d'une mixture Gaussienne est donnée par :

$$\begin{aligned} L_{x_1,\ldots,x_n}\left(\pi, \theta\right) &= \sum_{i=1} \log p\left(x_i \mid \pi, \theta\right) \\ &= \sum_{i=1}^{n} \log \left(\sum_{k=1} \pi_k \mathcal{N}\left(x_i; \mu_k, \Sigma_k\right) \right) \end{aligned}$$

Optimiser ceci n'est pas trivial, puisqu'on n'a pas d'expression analytique pour les estimateurs de π et θ.
En écrivant $Z_i \sim \mathcal{M}(1, \pi)$ et $X_i \mid Z_{ik} = 1 \sim \mathcal{N}(\mu_k, \Sigma_k)$, on a le modèle suivant :



2.2 Algorithme EM

Définition 2.2 En considérant l'échantillon dit complet des couples (X_i, Z_i), la log-vraisemblance *à données complètes* est donnée par :

$$\begin{aligned} L_{(x_i,z_i)_i}(\pi,\theta) &= \log p\left((x_i,z_i)_i \mid \pi,\theta\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\pi_k \mathcal{N}\left(x_i;\mu_k,\Sigma_k\right)\right). \end{aligned}$$

Proposition 2.2 Les estimateurs de π et θ maximisant la log-vraisemblance à données complètes sont :

- $\hat{\pi_k} = (1/n)\sum z_{ik}$
- $\hat{\mu_k} = (1/n_k)\sum z_{ik}x_i$
- $\hat{\Sigma_k} = (1/n_k)\sum z_{ik}(x_i - \hat{\mu_k})(x_i - \hat{\mu_k})^T$

Puisqu'on ne connaît pas les z_i et donc l'information de groupement qu'on cherche, on utilise l'algorithme EM (expectation maximisation), qui est une variante pondérée de l'algorithme des k-moyennes.

L'algorithme se fonde sur deux propriétés :

Lemme 2.1 Étant données les observations et les paramètres, les Z_i sont indépendants.

Lemme 2.2 Les probabilités $p\left(z_i \mid x_i,\pi,\theta\right) = \mathcal{M}(z_i,1,\tau_i)$ où :

$$\tau_{ik} = \frac{\pi_k \mathcal{N}(x_i,\mu_k,\Sigma_k)}{\sum_{l=1}^{K}\pi_l \mathcal{N}(x_i,\mu_l,\Sigma_l)}$$

τ_{ik} est la probabilité que l'observation i soit dans le groupe k.

On peut donc calculer la log-vraisemblance par rapport aux Z_i considérés comme variables aléatoires :

Proposition 2.3

$$\mathbb{E}_{(Z_i)_i}\left[L_{(x_i,Z_i)_i}(\pi,\theta)\right] = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\left(\pi_k\mathcal{N}\left(x_i,\mu_k,\Sigma_k\right)\right)$$

Proposition 2.4 Les estimateurs de π et θ maximisant la log-vraisemblance à données complètes sont :

- $\hat{\pi_k} = (1/n)\sum \tau 0_{ik}$
- $\hat{\mu_k} = (1/n_k)\sum \tau_{ik}x_i$
- $\hat{\Sigma_k} = (1/n_k)\sum \tau_{ik}(x_i - \hat{\mu_k})(x_i - \hat{\mu_k})^T$

On a alors l'algorithme suivant :

Algorithme 1 EM

On initialise les groupes, par exemple avec k-moyennes. Ensuite, tant que la log-vraisemblance change :

M Calculer $\hat{\pi}$ et $\hat{\theta}$ par rapport aux τ_i

E Calculer les τ_i par rapport à π et θ.

On vérifie aisément qu'une itération d'EM augmente la log-vraisemblance, mais on n'a pas de garantie, en général, de converger vers un maximum global.

Pour estimer le nombre K de groupes à partir des données, on applique l'algorithme pour différentes valeurs de K et on choisit un critère, basé sur M_K le nombre de paramètre libres :

- Le critère d'information bayésienne : $L_{x_i}(\hat{\pi}, \hat{\theta}) - (M_K/2) \log n$
- Le critère d'information d'Akaike : $L_{x_i}(\hat{\pi}, \hat{\theta}) - M_K$.

3 Régression Linéaire et Choix de Modèle

3.1 Régression Linéaire Bayésienne

Définition 3.1 Le *modèle de régression linéaire* est donné par $Y = X\beta + \varepsilon$ où :

- $Y \in \mathbb{R}^n$ est un vecteur d'élément y_i ;
- $X \in \mathcal{M}_{n,p}(\mathbb{R})$ est une matrice dont la i-ème ligne est ${}^{\mathrm{t}}x_i$;
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

On peut alors introduire une distribution *a priori* pour le vecteur de régression β :

$$p(\beta) = \mathcal{N}(\beta; 0, \frac{I_p}{\alpha})$$

avec $\alpha > 0$ fixé.

L'estimateur de vraisemblance maximale pour le vecteur de poids β est donc :

$$\hat{\beta} = ({}^{\mathrm{t}}XX)^{-1\,\mathrm{t}}XY$$

qui n'est calculable que si ${}^{\mathrm{t}}XX$ est de rang plein.

Proposition 3.1 Dans l'approche bayésienne, chercher un estimateur a posteriori maximal $\hat{\beta}_{\mathrm{MAP}}$ est équivalent à calculer l'estimateur de crête :

$$\begin{aligned}\hat{\beta}_{\mathrm{MAP}} &= \mathrm{argmax}_\beta \log p\left(\beta \,\middle|\, X, Y, \sigma^2\right) \\ &= \mathrm{argmin}_\beta \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}\end{aligned}$$

avec $\lambda = \alpha\sigma^2$. On obtient alors, si $\lambda = \alpha\sigma^2$ est suffisamment grand, une expression calculable pour $\hat{\beta}_{\mathrm{MAP}}$ en grande dimension.

$$\hat{\beta}_{\mathrm{MAP}} = ({}^{\mathrm{t}}XX + \alpha\sigma^2 I_p)^{-1\,\mathrm{t}}XY$$

Proposition 3.2 Dans l'approche bayésienne, la distribution postérieure du vecteur de régression est donc :

$$p\left(\beta \,\middle|\, X, Y, \sigma^2\right) = \mathcal{N}(\beta, m_n, S_n)$$

avec :

$$m_n = ({}^{\mathrm{t}}XX + \alpha\sigma^2 I_p)^{-1\,\mathrm{t}}XY$$

et :

$$S_n = \left(\frac{{}^{\mathrm{t}}XX}{\sigma^2} + \alpha I_p\right)^{-1}$$

3.2 EM Revisité

On veut maintenant estimer α en tant qu'(hyper)paramètre à partir de l'ensemble d'entraînement. On remplace donc la distribution a priori par $p(\beta \mid \alpha) = \mathcal{N}\left(\beta, 0, \frac{I_p}{\alpha}\right)$.

Ceci nous amène à l'algorithme suivant :

Algorithme 2 Procédure d'Évidence

procedure E(α, σ^2)
  $S_n \leftarrow \left(\frac{{}^tXX}{\sigma^2} + \alpha I_p\right)^{-1}$
  $m_n \leftarrow ({}^tXX + \alpha\sigma^2 I_p)^{-1}{}^tXY$
end procedure

procedure M(S_n, m_n)
  $\alpha \leftarrow p/\left(\mathrm{Tr}\left(S_n\right) + {m_n}^t m_n\right)$
  $\sigma^2 \leftarrow \frac{1}{n}\left\{\|Y - Xm_n\|^2 + \mathrm{Tr}({}^tXXS_n)\right\}$
end procedure

procedure EM
  Initialiser α, σ^2
  while $L(\beta)$ change **do**
    $E(\alpha, \sigma^2)$
    $M(S_n, m_n)$
  return (α, σ^2)
end procedure

3.3 Processus Gaussiens

On a :

- $Y \mid X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$
- $\beta \mid \alpha \sim \mathcal{N}(0, \frac{I_p}{\alpha})$

Ici β est une variable aléatoire latente et les hyperparamètres sont α et σ^2.

Proposition 3.3 On a alors, par composition de gaussiennes :

$$Y \mid X, \sigma^2, \alpha \sim \mathcal{N}(0, \frac{X{}^tX}{\alpha} + \sigma^2 I_n)$$

On peut directement optimiser pour trouver la vraisemblance maximale ci-dessus, mais cela demande $\mathcal{O}(n^3)$ opérations.

Définition 3.2 Un *processus gaussien* peut se construire comme :

$$Y \mid X, \sigma^2, \theta \sim \mathcal{N}(0, C_n)$$

avec $C_n = \left(k\left(x_i, x_j\right)\right)_{i,j} + \sigma^2 I_n$ pour un certain *noyau* k.

Définition 3.3 Le *noyau exponentiel quadratique* est donné par :

$$k(x_i, x_j) = \theta_0 \exp(-\frac{\theta_1}{2} \|x_i - x_j\|^2) + \theta_2 + \theta_3 {}^{\mathsf{t}}x_i x_j$$

Proposition 3.4 Le problème d'optimisation ci-dessus donné pour les processus gaussiens s'écrit :

$$(\hat{\theta}, \hat{\sigma}^2) = \text{argmax}_{\theta,\sigma^2} \log \mathcal{N}(Y, 0, C_n)$$

Ici, on a une complexité en $\mathcal{O}(n^3)$ pour la résolution.

Toutefois, ce modèle a un grand avantage : si on considère (X, Y) notre ensemble d'entraînement à n éléments, et on considère une nouvelle observation x_{n+1}, on peut écrire X_{n+1} en ajoutant x_{n+1} à X et Y_{n+1} ajoutant une ligne y_{n+1} à Y, qui va vérifier :

$$Y_{n+1} \mid X_{n+1}, \theta, \sigma^2 \sim \mathcal{N}(0, C_{n+1})$$

avec :

$$C_{n+1} = \begin{pmatrix} C_n & k \\ {}^{\mathsf{t}}k & c \end{pmatrix}$$

où $k_i = k(x_{n+1}, x_i) = k(x_i, x_{n+1})$ et $c = k(x_{n+1}, x_{n+1}) + \sigma^2$.

Proposition 3.5 On obtient alors, par la propriété gaussienne :

$$y_{n+1} \mid X_{n+1}, \theta, \sigma^2 \sim \mathcal{N}(\bar{m}, \bar{\sigma}^2)$$