TP Statistiques 4

Juhyun Park, Angie Pineda, Nicolas Brunel

8 avril 2022

Vraisemblance pour plusieurs paramètres

Dans la session précédente, nous avons considéré un modèle simple à un seul paramètre. La plupart des modèles statistiques sont plus compliqués, impliquant souvent de nombreux paramètres inconnus. Il est évident que nous devons développer des méthodes plus flexibles et plus puissantes pour traiter la fonction de vraisemblance.

La Loi Normal

1. Soit $X_1, \ldots, X_n$ un échantillon de n variables indépendentes de loi de Normal(μ, σ^2) où $\theta = (\mu, \sigma)$ est inconnue. Simuler un échantillon i.i.d de taille $n = 25$ avec $\theta_0 = (0, 1)$. Présentez l'histogramme des données simulées. Choisir quatre paramètres candidats, disons, θ_0 (vrai) $\theta_1, \theta_2, \theta_3$. Comparer l'histogramme avec les densités candidates. Que remarquez-vous?
2. Ecrire la log vraisemblance. Générez une fonction de log-vraisemblance avec les arguments (θ, x), qui donne la log vraisemblance d'un échantillon pour une valeur donnée de $\theta = (\mu, \sigma)$ et les données $x = (x_1, \ldots, x_n)$. Calculez la log-vraisemblance des échantillons que vous avez générés, en faissant varier un paramètre à la fois. Présentez graphiquement la surface de log-vraisemblance que vous calculez, utilisant, par exemple, `contour`, `image` ou `persp`. Markez le point maximum. Que remarquez-vous?
3. Variez l'échantillon. Présentez l'histogramme des données simulées et la surface de vraisemblance correspondante. Que remarquez-vous?
4. Répétez pour les tailles d'échantillon croissantes $n = 10, 25, 50, 100$. Commentez.

Maximum de vraisemblance pour plusieurs paramètres

Pour des lois à plusieurs paramètres, résoudre un problème numérique nécessite de trouver un maximum sur une surface. Ceci pourrait être résolu en utilisant des solveurs plus compliqués que `optimize`. Nous utiliserons la fonction `optim` qui est la routine d'optimisation à usage général disponible dans R.

```
?optim
```

En utilisant `optim`, notez que cette routine est par défaut une routine de **minimisation**. Comme vous pouvez le voir, cette fonction peut implémenter diverses méthodes d'optimisation.

Les routines de maximisation, ou de manière équivalente de minimisation, visent à trouver la position $\hat{\theta}$ qui maximise la fonction. Ces techniques sont généralement appelées *quasi-Newton*. Dans les routines quasi-Newton, le paramètre est mis à jour par des pas successifs (de longueur et de direction variables) dans l'espace des paramètres jusqu'à ce que la convergence soit considérée comme atteinte. Les étapes sont généralement déterminées par le calcul interne des gradients locaux et de la courbure (comme avec Newton-Raphson) mais un algorithme numérique intelligent est utilisé pour éviter la nécessité d'une détermination numérique de ceux-ci après chaque étape. Il existe différents niveaux de complication ou d'effort requis pour l'utilisation du ces routines:

- aucune information autre que la fonction,

- seulement la fonction et ses premières dérivées données explicitement,
- la fonction, ses premières et secondes dérivées (hessian) donné explicitement.

Dans la plupart des cas, la première méthode est généralement tout à fait adéquate, plus rapide à utiliser et il y a moins de marge d'erreur car aucun dérivé analytique ne doit être trouvé ou programmé.

Parmi les arguments, la **method** spécifie la méthode de minimisation de la fonction. La méthode *Nelder-Mead* (défaut) est la plus lente, mais la plus robuste; et peut faire face à discontinuités dans la surface de vraisemblance.

La seule méthode qui peut faire face aux contraintes de paramètres est *L-BFGS-B*.

Gérer les contraintes sur les paramètres

Il y a souvent des contraintes sur les paramètres, ce qui pourrait affecter les performances des algorithmes numériques.

Avec la méthode de *L-BFGS-B* dans la fonction `optim`, trois types de conditions aux limites peuvent être traités:

1). **Plages marginales de paramètres**: par exemple, $X_i \sim N(\mu, \sigma^2)$ pour $i = 1, \ldots, n$, alors pour $\theta = (\mu, \sigma)$ nous avons $\Omega = (-\infty, \infty) \times (0, \infty)$. Les routines de minimisation essaieront souvent les valeurs limites pour aider à apprendre la fonction de vraisemblance, donc on essayera $\sigma = 0$ disons, pour lequel la vraisemblance l'évaluation explose. Pour éviter de tels problèmes, il est utile de contracter le paramètre espace pour être $\Omega = [-10^6 < \mu < 10^6] \times [10^{-4} < \sigma < 10^6]$ dire.

2). **Contraintes entre paramètres**: par exemple, (a) $X_i \sim \text{Gamma}(\alpha, \beta)$. Nous devons adapter un modèle avec une contrainte sur l'espérance $E(X) \leq 10$ (par exemple). Cette contrainte correspond à $\alpha \leq 10\beta$. (b) $\mathbf{X}_i \sim \mathcal{N}(\mu, \Sigma)$ où $\mathbf{X} = (X(t_1), \ldots, X(t_m))$ est un processus gaussien observé à $(t_1, \ldots, t_m)$. Lors de l'ajustement d'un modèle pour Σ, la matrice de variance-covariance doit être définie positive pour tous les points possibles dans l'espace des paramètres. Cela impose des contraintes sur les paramètres. Ces contraintes peuvent généralement être gérées soit par (i) en utilisant une routine quasi Newton plus sophistiquée, ou (ii) en mettant des contraintes dans la routine de vraisemblance - par exemple la loi Gamma ci-dessus définit la log-vraisemblance des paramètres $\alpha > 10\beta$ pour être $-\infty$, ce qui est réalisé en pratique par prenant $-\log L(\theta) = 10^6$ pour θ dans cette plage.

3). **Contraintes issues des données**: par exemple, $X_i \sim U(0, \theta), i = 1, \ldots, n$. Ici la seule contrainte pour θ qui peut être spécifié avant les données est que $\theta > 0$. Toutefois on sait que $L(\theta) = 0$ quand $\theta < \max(x_i)$. Ces informations doivent être intégrées dans la vraisemblance/log-vraisemblance fonction directement. Dans cet exemple, nous devons définir $-\log L(\theta) = 10^6$, par exemple, c'est-à-dire une valeur beaucoup plus petite, vraisemblablement, que celles de l'espace des paramètres valides.

Dans tous ces cas, vous devez prendre soin de vérifier si l'estimateur du maximum de vraisemblance est sur la frontière. Dans de nombreux cas, s'il se trouve à la limite, il y aura des problèmes numériques pour dériver le variance/erreurs standard des paramètres.

5. En utilisant la fonction `optim`, trouver la valeur de θ la plus probable pour l'échantillon de normal. En fait, l'estimateur est defini explicitement pour le cas normal. Verifiez la solution numerique avec la solution analytique.

6. Testez avec des échantillons variés et trace l'histogramme de l'estimateur. Variez la taille de l'échantillons $n = 10, 25, 50, 100$ et comparez l'écart entre la valeur théorique attendue et la valeur obtenue. Que remqrauez-vous?

La Loi Gamma

7. Soit $X_1, \ldots, X_n$ un échantillon de n variables indépendentes de loi de Gamma(α, β) où $\theta = (\alpha, \beta)$ est inconnue. Simuler un échantillon i.i.d de taille $n = 25$ avec $\theta_0 = (3, 1)$. Présentez l'histogramme des données simulées. Choisir quatre paramètres candidats, disons, θ_0 (vrai) $\theta_1, \theta_2, \theta_3$. Comparer l'histogramme avec les densités candidates. Que remarquez-vous?

8. Ecrire la log vraisemblance. Générez une fonction de log-vraisemblance avec les arguments (θ, x), qui donne la log vraisemblance d'un échantillon pour une valeur donnée de $\theta = (\alpha, \beta)$ et les donnéé $x = (x_1, \ldots, x_n)$. Calculez la log-vraisemblance des échantillons que vous avez générés, en faissant varier un paramètre à la fois. Présentez graphiquement la surface de log-vraisemblance que vous calculez et markez le point maximum. Que remarquez-vous? Est-ce qu'il y a quelque chose de notable par rapport au cas normal?

9. Variez l'échantillon. Présentez l'histogramme des données simulées et la surface de vraisemblance correspondante. Que remarquez-vous?

10. Répétez pour les tailles d'échantillon croissantes $n = 10, 25, 50, 100$. Commentez.

11. En utilisant la fonction `optim`, trouver la valeur de θ la plus probable pour l'échantillon de normal. En fait, l'estimateur est defini explicitement pour le cas normal. Verifiez la solution numerique avec la solution analytique.

12. Testez avec des échantillons variés et trace l'histogram de l'estimateur. Variez la taille de l'échantillons $n = 10, 25, 50, 100$. et comparez l'écart entre la valeur théorique attendue et la valeur obtenue. Que remqrauez-vous?

Application aux données sur l'ozone

13. Utilisez les données sur l'ozone ("summer_ozone.csv", "winter_ozone.csv") pour trouver l'estimateur de maximum de vraisemblance pour chaque site à chaque saison si on considére que c'est (i) une loi normale et (ii) une loi log normale. Comparez les résultats. Quel modèle préférez-vous et pourquoi?