Bayesian matrix factorization for outlier detection: an application in population genetics

Michael G.B. Blum[1], Nicolas Duforet-Frebourg[1]

Bayesian Young Statisticians Meeting (BAYSM), Milan June, 5-6, 2013
Paper no. 13

[1] Laboratoire TIMC-IMAG UMR 5525, Centre National de la Recherche Scientifique,
Université Joseph Fourier, Grenoble, France
`nicolas.duforet@imag.fr`
`michael.blum@imag.fr`

Abstract

We present a new Bayesian hierarchical model based on matrix factorization for detecting outliers in high-dimensional data. Outliers are explicitly modeled using both a shift in mean and variance inflation approach. The Bayesian framework provides intrinsic probabilities of being an outlier for each element in the sample. Posterior replicates of the parameters are simulated using a MCMC algorithm. In population genetics where many genetic markers are typed in different populations, we show that this model can be used to detect genes targeted by Darwinian selection.

Keywords: Bayesian matrix factorization; factor model; principal component analysis; outlier detection; population genetics.

1 Introduction

Matrix factorization aims at decomposing a high-dimensional $n \times p$ data matrix into a product of two lower rank K matrices called the factor and loading matrices [3]. Matrix factorization provides a useful framework to model outliers in the lower-dimensional space generated by the low rank approximation [2]. Detecting outliers in high dimensional data sets is of interest in population genetics in order to detect genes under selective pressures [1]. The proposed approach provides an intrinsic probability of being an outlier so that we can estimate false discovery rate (FDR) and q-values, which are two important quantities in whole-genome scans [5].

We provide a MCMC algorithm to sample replicates from the posterior distribution and we show how the method can detect genes under selection in population genetics data.

2 Bayesian matrix factorization for outlier detection

2.1 Model

The probabilistic model of matrix factorization—also known as factor or probabilistic PCA model—for a design $n \times p$ matrix Y relies on a product between a factor matrix F and a loading matrix Λ,

$$Y = F\Lambda + \epsilon, \tag{1}$$

where F is a $n \times K$ matrix, Λ is a $K \times p$ matrix, and ϵ is a $n \times p$ residual matrix where each row $\epsilon_i \sim \mathcal{N}(0_p, \sigma^2 I_p)$. Here, we choose a Gaussian prior for Λ

$$p(\Lambda|\sigma_\Lambda) = \Pi_{j=1}^{p} \mathcal{N}(\Lambda_j; 0_K, \sigma_\Lambda^2 I_K). \tag{2}$$

To specify the prior of F, we explicitly model outliers using the shift-in-mean approach [4] for one of the K factors of the low-rank approximation

$$p(F|A, Z, \Sigma_F) = \Pi_{i=1}^{n} \mathcal{N}(F_i; 0_K + A_i^{(Z_i)}, \Sigma_F). \tag{3}$$

where Σ_F is a diagonal matrix with values $\sigma_{F_k}^2$. We specify improper priors for variances $p(\sigma_\Lambda^2) \propto \frac{1}{\sigma_\Lambda^2}$, and $p(\sigma_{F_k}^2) \propto \frac{1}{\sigma_{F_k}^2}$. Shift vectors A_i's are zero valued vectors with non-zero component at index Z_i. For $i = 1, \dots, n$, Z_i is an integer between 1 and $K+1$, indicating that the i^{th} line is either an outlier for the factor Z_i if $Z_i < K + 1$ or not an outlier if $Z_i = K + 1$. We add priors for A and Z such as

$$p(A|\tau, \Sigma_F) = \Pi_{i=1}^{n} \mathcal{N}(A_i; 0_K, \tau^2 \Sigma_F). \tag{4}$$

$$p(Z_i = k) = \pi_k = \left\{ \begin{array}{l} \alpha/K \text{ if } k < K+1 \\ 1 - \alpha \text{ if } k = K+1 \end{array} \right. \tag{5}$$

where α is the expected proportion of outliers that is set *a priori*, and τ is the variance-inflation parameter with $p(\tau) = \mathcal{U}(1, 10)$.

2.2 Posterior inference and algorithm

To obtain replicates from the posterior $p(Z, A|Y)$, we use Gibbs updating steps based on the conditional distribution of (Z_i, A_i) provided below

$$p(Z_i = k, A_i|F_i, \Sigma_F, \tau^2) = p(Z_i = k|F_i, \Sigma_F, \tau^2) p(A_i|Z_i = k, F_i, \Sigma_F, \tau^2) \tag{6}$$

where

$$p(Z_i = k|F_i, \Sigma_F, \tau^2) \propto \pi_k \sqrt{\sigma_{F_k}^2} e^{\frac{\tau^2}{(\tau^2+1)} \frac{F_{i,k}^2}{\sigma_{F_k}^2}} \tag{7}$$

$$p(A_{i,k}|Z_i = k, F_i, \Sigma_F, \tau^2) \propto \mathcal{N}(\frac{\tau^2}{\tau^2+1}F_{i,k}, \frac{\tau^2}{\tau^2+1}\sigma^2_{F_k}) \text{ if } k < K+1. \quad (8)$$

Other parameters have more usual conditional distributions that are also useful when performing Bayesian linear regression. Samples are simulated using the MCMC algorithm provided in Table 1.

```
• Setup values of   $\alpha, K.$
• Initialize        $\sigma, \sigma_\Lambda, \Sigma_F, \Lambda, F, A, Z, \tau^2.$
• for $s = 1..ns$ do:
                    $j = 1..p, \Lambda_j^{(s)} \leftarrow \mathcal{N}((\frac{1}{\sigma_\Lambda^{2(s-1)}}I_K + \frac{1}{\sigma^{2(s-1)}}(F^{(s-1)})^t F^{(s-1)})^{-1}\frac{1}{\sigma^{2(s-1)}}F^{(s-1)}Y_j,$
                    $(\frac{1}{\sigma^{2(s-1)}}F^{(s-1)t}F^{(s-1)} + \frac{1}{\sigma_\Lambda^{2(s-1)}}I_K)^{-1})$
                    $\sigma_\Lambda^{(s)} \leftarrow IG(\frac{Kp}{2}, \frac{1}{2}\sum_{i=1}^K \Lambda_i^{(s)}\Lambda_i^{(s)t})$
                    $i = 1..n, Z_i^{(s)} \leftarrow sample(Z_i^{(s)}, p(Z_i = k|\pi, F_i^{(s-1)}, \Sigma_F^{(s-1)}, \tau^{2(s-1)}))$
                    $i = 1..n, \text{if } Z_i^{(s)} < K+1 \text{ then}, A_{i,Z_i}^{(s)} \leftarrow \mathcal{N}(\frac{\tau^{2(s-1)}}{\tau^{2(s-1)}+1}F_{i,Z_i}^{(s-1)}, \frac{\tau^{2(s-1)}}{\tau^{2(s-1)}+1}\sigma_{F_{Z_i}}^{2(s-1)})$
                    $i = 1..n, F_i^{(s)} \leftarrow \mathcal{N}((\Sigma_F^{-1(s-1)} + \frac{1}{\sigma^{2(s-1)}}\Lambda^{(s)}\Lambda^{t(s)})^{-1}$
                    $(\Sigma_F^{-1(s-1)}A_{i,Z_i}^{(s)} + \frac{1}{\sigma^{2(s-1)}}\Lambda^{(s)}Y), (\Sigma_F^{-1(s-1)} + \frac{1}{\sigma^{2(s-1)}}\Lambda^{(s)}\Lambda^{t(s-1)})^{-1})$
                    $i = 1..K, \sigma_{F_i}^{(s)} \leftarrow IG(\frac{n}{2}, \frac{1}{2}\sum_{j=1}^n (F_{ij}^{(s)} - A_{ij}^{(s)})(F_{ij}^{(s)} - A_{ij}^{(s)})^t)$
                    $\sigma^{(s)} \leftarrow IG(\frac{np}{2}, \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^p |Y_{i,j} - F_i^{(s)}\Lambda_j^{(s)}|))$
                    Metropolis-Hastings step: $\tau^{2*} \leftarrow \mathcal{N}(\tau^{2(s-1)}, .5)$
```

Table 1: MCMC algorithm of Bayesian Matrix Factorization for detecting outliers.

3 Results

To illustrate the potential of the method, we simulate population genetics data where the outliers corresponds to the markers located in genomic regions under Darwinian selection. The data contain 400 individuals from 4 populations that split according to a tree model (see Panel A in Figure 1), and are typed at 200 genetic markers called SNPs, among which 12 are under various selective pressures in one of the 4 populations. Posterior probabilities to be outliers are enriched for genes targeted by selection (see Panel B in Figure 1), and a Precision-Recall (see Panel C in Figure 1) curve can be used to evaluate the property of the method under various evolutionary scenario.

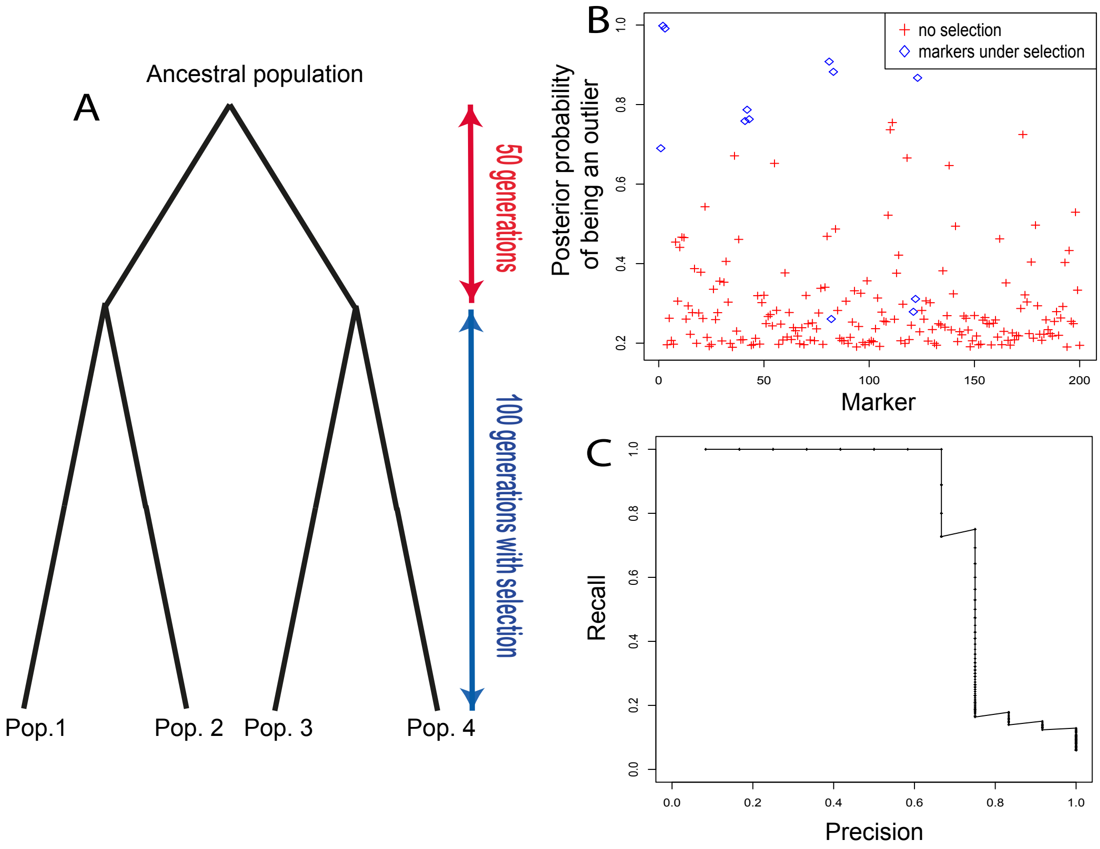

Figure 1: Panel A: Population divergence model used to simulate population genetics data. Panel B: Posterior Probability of being outlier. Panel C: Precision Recall curve.

4 Conclusions

We introduced a hierarchical Bayesian model of matrix factorization for detecting outliers in high-dimensional data. The probabilistic model provide probabilities for each observation to be outlier and indicates the direction or factor under which the observation is atypical. We showed that this approach can be used to detect genes targeted by selection in population genetics.

References

[1] M.A. Beaumont, D.J. Balding. **Identifying adaptive genetic divergence among populations from genome scans**. *Molecular ecology*; (2004); 13(4); 969-980.

[2] W. Polasek. **Factor analysis and outliers: A Bayesian approach.** *Wirtschaftswissenschaftliches Zentrum der Universitt Basel.* 1997.

[3] R. Salakhutdinov, A. Mnih. **Bayesian probabilistic matrix factorization using Markov chain Monte Carlo**. *In Proceedings of the 25th international conference on Machine learning*; (2008); (pp. 880-887). ACM.

[4] I. Verdinelli, L. Wasserman. **Bayesian analysis of outlier problems using the Gibbs sampler**. *Statistics and Computing*; (1991); 1(2); 105-117.

[5] J. Wakefield. **A Bayesian measure of the probability of false discovery in genetic epidemiology studies**. *American journal of human genetics*; (2007); 81(2); 208.