High-dimensional Bayes

Debdeep Pati
Florida State University

November 6, 2014

Background on factor models

- Massive dimensional vector of candidate predictors encountered in many application areas.
- Factor models provide a convenient framework for dimension reduction in large p, small n applications (West, 2003; Lucas et al., 2006; Carvalho et al., 2008).
- Explain dependence among high dimensional observations through fewer number of underlying factors.

Factor Models

- Dependence in the high dimension observations explained partially through shared dependence on some latent factors

$$y_i = \Lambda\eta_i + \epsilon_i, \quad cov(\epsilon_i) = \Omega$$

- Λ: Factor loadings, η_i: factor corresponding to the ith observation
- ϵ_i are idiosyncratic noise.

Principal component Analysis

- PCA is a orthogonal linear transformation to the data
- Transforms the data to a new coordinate system
- Greatest variance direction is the direction of the first coordinate
- Centered $n \times p$ data matrix X
- $Y^T Y = W \Lambda W^T$,
- Λ diagonal matrix of eigen values, columns of W corresponding eigen vectors
- $T = YW$ are the principal components

Motivating applications: High dimensional regression

- Develop accurate predictive models for health outcomes based on high-dimensional biomarkers.
- $z_i \in \Re$ some continuous health outcome. $x_i \in \Re^{p-1}$ vector of candidate predictors.
- Sparse factor model for $y_i = (z_i, x_i) \in \Re^p$ jointly.
- Regularized estimation of joint covariance matrix.
- Prediction and variable selection based on induced conditional $E(z_i \mid x_i)$.

Motivating applications: Large covariance matrix estimation

- Interest in modeling $Cov(y_i)$
- Factor models provide a natural approach
- $Cov(y_i) = \Lambda\Lambda' + \Omega$
- Low rank + sparse decomposition

Motivating applications: Subspace estimation

- Interest in learning the low dimensional subspace on which y_is lie
- Estimate Λ
- Considerably harder problem due to identifiability issues
- Can make Λ semi-orthogonal matrix
- Leads to Probabilistic Principal Component Analysis (PPCA)
- Still not enough for subspace estimation

Gaussian Linear Factor Models

- Jointly model y_i's after normalizing as

$$y_i = \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma), \quad i = 1, \ldots, n,$$

- Λ is a $p \times k$ factor loadings matrix, $\eta_i \sim N_k(0, I_k)$ are latent factors and ϵ_i idiosyncratic error with $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.
- Marginalizing over the latent factors, $y_i \sim N_p(0, \Omega)$ with $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$.

Bayesian factor models - recent developments

- Variable selection-type mixture prior on loadings (Lucas et al., 2006; Carvalho et al., 2008).
- Recent work on latent feature models using the Indian buffet process (Griffiths & Ghahramani, 2006; Thibaux & Jordan, 2007).
- Weighted versions have found applications in factor analysis (Knowles & Ghahramani, 2007; Meeds et al., 2007; Rai & Daumé, 2009).
- Parameter expansion to induce heavy-tailed default prior on the loadings (Ghosh & Dunson, 2009).

Focus on Regression and Covariance matrix estimation

- Identifiability of the loadings not necessary in many applications
- Variable selection-type mixture priors need many one-at-a-time updates – mixes slowly and computationally challenging.
- Heavy-tailed shrinkage prior on loadings instead, loadings increasingly shrunk to zero with column index.
- Allows block updating of loadings and selection of truncation level.

Some notations

- Θ_Λ to denote the collection of matrices Λ with p rows and infinitely many columns such that $\Lambda\Lambda^{\mathrm{T}}$ is a $p \times p$ matrix with all entries finite.

$$\Theta_\Lambda = \left\{ \Lambda = (\lambda_{jh}),\, j = 1, \ldots, p,\ \ h = 1 \ldots, \infty,\ \max_{1 \le j \le p} \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \right\}$$

The MGPS prior (Bhattacharya & Dunson, 2011 (Biometrika)

- Proposed multiplicative gamma process shrinkage (MGPS) prior on the loadings is given by

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}),\ \phi_{jh} \sim \mathcal{G}(\nu/2, \nu/2),$$

$$\tau_h = \prod_{l=1}^{h} \delta_l,\ \delta_1 \sim \mathcal{G}(a_1, 1),\ \delta_l \sim \mathcal{G}(a_2, 1),\ l \geq 2,$$

- τ_h is a global shrinkage parameter for the hth column, stochastically increasing under the restriction $a_2 > 1$.
- ϕ_{jh}'s are local shrinkage parameters for the elements in the hth column, avoid over-shrinking the non-zero loadings in later columns.

Truncation approximation error

- For computational purposes, approximate the infinite loadings matrix with a finite matrix having few columns relative to p.
- We obtain theoretical bounds on the truncation approximation error.
- Let $(\Lambda, \Sigma) \sim \Pi_\Lambda \otimes \Pi_\Sigma$ and $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$. We can approximate Ω by $\Omega_T = \Lambda_T \Lambda_T^{\mathrm{T}} + \Sigma$.

Theorem

If $a_2 > 2$, then for any $\epsilon > 0$,

$$pr\big\{d_\infty\big(\Omega, \Omega_T\big) > \epsilon\big\} < \frac{6pb}{\epsilon(1-a)} a^T \ \textit{for} \ \ T > \frac{\log\{6pb/\epsilon(1-a)\}}{\log(1/a)},$$

where $b = E\big(\delta_1^{-1}\big)$ and $a = E\big(\delta_2^{-1}\big)$.

Choice of the truncation level

- Truncate the loadings matrix to have $k^* << p$ columns. Posterior samples from approximated conditional posterior.
- How to chose an appropriate level of truncation?
- Redundant factors – correspond to columns of loadings whose all elements are less than ϵ in magnitude.
- Effective factors – all non-redundant factors.

A possible approach

- Start with a conservative guess $\tilde{k}$ of k^*.
- At the tth iteration of the Gibbs sampler, define $m^{(t)}$ to be the number of redundant columns in $\Lambda_{\tilde{k}}$, whose all elements are less than ϵ in magnitude($\epsilon = 10^{-4}$ used as a default)
- Usual shrinkage priors on the loadings exhibit the phenomenon of factor splitting.
- Our approach avoids this problem by shrinking increasingly in later columns.
- Define $k^{*(t)} = \tilde{k} - m^{(t)}$ to be the effective number of factors at iteration t.

Adaptive Gibbs sampler

- Adapt the number of factors as the sampler progresses – avoids specifying over-conservative initial guess.
- Designed to satisfy the diminishing adaptation condition of Roberts & Rosenthal (2007). Discard redundant columns if $m^{(t)} > 0$, otherwise add a new column with additional parameters drawn from the prior.
- Let $\tilde{k}^{(t)}$ be the truncation level at the tth iteration and $k^{*(t)} = \tilde{k}^{(t)} - m^{(t)}$ the effective number of factors.
- Estimate k^* by the mode or median of the samples $\{k^{*(t)}\}_{t=B+1}^{N}$.

Covariance matrix estimation

- Set $\Omega^{(t)} = \Lambda^{(t)}_{\tilde{k}^{(t)}} \Lambda^{(t)\prime}_{\tilde{k}^{(t)}} + \Sigma^{(t)}$.
- $\{\Omega^{(t)}\}_{t=B+1}^{N}$ represent draws from the approximated marginal posterior distribution of Ω given $y_i, i = 1, \ldots, n$.

Regression Coefficient Estimation

- Recall, after marginalizing out latent factors, $y_i \sim N_p(0, \Omega)$ with $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$.
- $E(z_i \mid x_i) = x_i^{\mathrm{T}}\beta$, with $\beta = \Omega_{xx}^{-1}\,\Omega_{zx}$, true regression coefficients of z on x.
- Set $\beta^{(t)} = \{\Omega_{xx}^{(t)}\}^{-1}\,\Omega_{zx}^{(t)}$, where $\Omega_{xx}^{(t)} = \Lambda_x^{(t)}\Lambda_x^{(t)\,\mathrm{T}} + \Sigma_{xx}^{(t)}$ denote posterior samples at the tth iteration.
- Computation involves inverting $\tilde{k}^{(t)} \times \tilde{k}^{(t)}$ matrices at tth iteration.
- Let $\hat{\beta}$ denote the posterior mean of β. The proposed formulation retains the non-zero elements of β while heavily shrinks the rest toward zero.