Bayesian methods for high dimensional models: Convergence issues and computational challenges

Subhashis Ghosal,
North Carolina State University

van Dantzig Seminar,
University of Amsterdam
June 3, 2013
Based on collaborations with
Sayantan Banerjee, Weining Shen and S. McKay Curtis

Some High Dimensional Statistical Models

- Normal mean: $Y_i \overset{\text{ind}}{\sim} N(\theta_i, \sigma^2)$, $i = 1, \ldots, n$.
- Linear regression: $Y_i = \beta' X_i + \varepsilon_i$, independent errors (possibly normal) with variance σ^2, $i = 1, \ldots, n$, $\beta \in \mathbb{R}^p$, possibly $p \gg n$, can even be exponential in n.
- Generalized linear model: $Y_i \overset{\text{ind}}{\sim} \text{ExpFamily}(g(\beta' X_i))$, $i = 1, \ldots, n$, g some link function, $\beta \in \mathbb{R}^p$, possibly $p \gg n$.
- Normal covariance (or precision): $X_i \overset{\text{iid}}{\sim} N_p(0, \Sigma)$, $i = 1, \ldots, n$, possibly $p \gg n$.
- Exponential family: $X_i \overset{\text{iid}}{\sim} \text{ExpFamily}(\theta)$, $\theta \in \mathbb{R}^p$, possibly $p \gg n$.
- Nonparametric additive regression: $Y_i \overset{\text{ind}}{\sim} N(\sum_{j=1}^{p} f_j(X_{ij}), \sigma^2)$, $i = 1, \ldots, n$, $f_1, \ldots, f_p$ smooth functions acting on p co–ordinates of covariate X, possibly $p \gg n$.
- Nonparametric density regression: $Y_i | X_i \overset{\text{ind}}{\sim} f(\cdot \mid X_i)$, f smooth, X_i's p-dimensional, possibly $p \gg n$.

Sparsity

- Sparsity — Only a few of stated relations are non-trivial.
- An essential low dimensional structure, often present in high dimensional models, making inference possible.
  - Normal mean: Only $r \ll n$ means are non-zero.
  - Linear regression: Only $r \ll \min(p, n)$ coefficients are non-zero.
  - Normal covariance (or precision):
    - (Nearly) banding structure: Total contribution of off-diagonal elements outside a band is small;
    - Graphical model structure: Off-diagonal elements are non-zero only if the the corresponding edges are connected.
  - Nonparametric additive regression: Only $r \ll \min(p, n)$ functions are non-zero.
  - Nonparametric density regression: Only $r \ll \min(p, n)$ covariates actually influence the conditional density.

More Settings of Sparsity

- Estimating missing entries of a large matrix: A large matrix, whose entries are observed with errors, have many entries missing. Assume that the $p \times p$ matrix is expressible as $A + BC$, where A is sparse (meaning most entries are zero, like a diagonal or a thinly banded matrix) and B and C are low rank matrices (line $p \times r$ and $r \times p$, where $r \ll p$, say $r = 1$).
- Clustering: $X_i \overset{\text{ind}}{\sim} N(\theta_i, \sigma^2)$, many θ_i's are tied with each other to form $r \ll n$ groups. Tieing patterns and cluster means $\xi_1, \ldots, \xi_r$, as well as r, are unknown.

Oracle

- If sparsity structure is known, then inference reduces to low dimensional analysis, and hence optimal procedures are clear. For instance, in the normal mean model, if we knew which θ_i's are non-zero, we just estimate them incurring estimation error $r\sigma^2$ rather than $n\sigma^2$.
- The goal is to match the performance of the oracle within a small extra cost (which may come in the form of additive and/or multiplicative constant, and sometimes an additional log factor). For instance, in the sequence model, unless the oracle is known, a logarithmic factor is unavoidable.
- If signals are sufficiently strong, one also likes to discover the true sparsity structure up to small error (for instance, one likes to conclude, with probability tending to one, the estimated sparsity agrees with the true sparsity).

Classical Procedures for High Dimensional Data

- The most famous classical procedure for detecting sparsity in linear regression is the Lasso [Tibshirani (1996)]. It imposes an ℓ_1-penalty to set certain coefficients to zero, thus leading to a sparse regression.
- Recent book Bühlman and van de Geer (2011) studies theoretical aspects of Lasso and related methods thoroughly.
- Covariance estimation under (nearly) banding structure was developed by Bickel and Levina (2008) and others.
- To estimate a covariance matrix under the the graphical model setting can be done by imposing ℓ_1-penalty on entries, leading to the so called graphical Lasso.

Bayesian Procedures for High Dimensional Data

- We are interested in Bayesian procedures for high dimensional data. Bayesian procedures also give assessments of model uncertainty and lead to more natural approach to prediction.
- Sparsity is easily incorporated in a prior, for instance, by putting a Dirac point mass at zero.
- How does one approach posterior computation when dimension is very high? Changing dimension suggests Reversible Jump MCMC, but does not work at this scale.
- What can one say about concentration of the posterior distribution near the truth? Does it (nearly) match the oracle?
- Does a sparse version of the Bernstein-von Mises theorem hold, i.e., the posterior is asymptotically the product of normal of the oracle dimension and Dirac masses at zero?

Behavior of Posterior in High Dimension without Sparsity

- For generalized linear regression, linear (possibly non-normal) regression and exponential family models, Ghosal (1997, 1999, 2000) respectively obtained convergence rates and Bernstein-von Mises theorem for the posterior distribution for $p \to \infty$ without sparsity, but needed $p \ll n$.
- Influenced by the works of Portnoy (1986, 1986, 1988) and Haberman (1977) for similar results on MLE.
- Will be interesting to investigate sparse Bernstein-von Mises theorems so that $p \gg n$ will be allowed.

Normal Mean Model

- Castillo and van der Vaart (2012) considered mixture of point mass and heavy tailed prior, and showed that with high posterior probability $\|\theta - \theta_0\|^2$ is of the order $r \log(n/r)$ (agreeing with the minimax rate), and also that the support of the θ has cardinality of the order r. This can be considered as a full Bayesian analog of the empirical Bayes approach of Johnstone and Silverman (2004).
- They also have a smart computational strategy evaluating model probabilities as coefficients of a certain polynomial, but is very tied to the normal mean model.
- Babenko and Belitser (2010) considered an oracle formulation and showed that $\|\theta - \theta_0\|^2$ if of the order of the "oracle risk" with high posterior probability.

Generalized Linear Model

- Jiang (2007) studied posterior convergence rates for generalized linear regression under sparsity where $\log p = O(n^{\alpha})$, $\alpha < 1$ and obtained the rate $n^{-(1-\alpha)/2}$.

Computation: Linear Regression

- Bayesian Lasso [Park and Casella (2008)]: Linear regression using Laplace prior and MCMC. No point mass.
- Stochastic Search Variable Selection [Geroge and McCullagh (1993)] using spike and slab prior — really a low dimensional affair.
- Laplace approximation technique [Yuan and Lin (2005)]:
  - Posterior probabilities of various models are given by integrals of likelihood (a product of n functions) and the prior, which is taken as independent Laplace on non-zero coefficients. Use the fact that the posterior mode is Lasso restricted to the model. Expand the log-likelihood around the posterior mode and use Laplace approximation.
  - Works only for "regular models", for which no estimated coefficient is zero, i.e., only subsets of Lasso selection.
  - Every "non-regular model" is dominated by the corresponding regular model in terms of model posterior probability.

Nonparametric Additive Regression

- Use Yuan and Lin's (2005) idea of Laplace approximation to compute model posterior probabilities.
- Expand each function in a basis: $f_j(x_j) = \sum_{l=1}^{m_j} \beta_{j,l}\psi_{j,l}(x_j)$.
- The corresponding group of coefficients are given independent multivariate Laplace prior along with Dirac mass at zero.

$$\begin{aligned} p(\beta_j|\gamma) &= (1-\gamma_j)\mathbb{1}(\beta_j = 0) \\ &\quad + \gamma_j \frac{\Gamma(m_j/2)}{2\pi^{m_j/2}\Gamma(m_j)} \left(\frac{\lambda}{2\sigma^2}\right)^{m_j} \exp\{-\frac{\lambda}{2\sigma^2}\|\beta_j\|\}. \end{aligned}$$

  Also $p(\gamma) \propto d_\gamma q^{|\gamma|}(1-q)^{p-|\gamma|}$.
- The posterior mode now corresponds to the group Lasso [Yuan and Lin (2006)], restricted to the model. Always the case for additive penalty with minimum zero at zero.

Nonparametric Additive Regression (contd.)

- For Laplace approximation, need to calculate the Hessian of the log-posterior at the posterior model in model γ: $\sigma^{-2}(2\Psi_\gamma^T\Psi_\gamma + \lambda A_\gamma)$, where Ψ_γ is the data matrix considering the expansion and A_γ is a block-diagonal matrix coming from the Hessian of the log-prior (present due to the multivariate nature).
- Model posterior probabilities for all regular models are approximately proportional to

$$d_\gamma(q\lambda/2(1-q))^{|\gamma|}\prod_{j\in J_\gamma}(\Gamma(m_j/2)/\Gamma(m_j))$$
$$\times \det(\Psi_\gamma^T\Psi_\gamma + \frac{\lambda}{2}A_\gamma)^{-1/2}$$
$$\times \exp\{-[\|Y-\Psi_\gamma\hat{\beta}_\gamma\|^2 + \lambda\sum_{j\in J_\gamma}\|\hat{\beta}_j\|]/(2\sigma^2)\}.$$

Nonparametric Additive Regression (contd.)

- If $q < 1/2$, non-regular models are dominated by regular models, so search for high posterior probability models may be restricted to regular models only.
- Consistency of group Lasso selection means that correct model is regular with probability tending to one.
- Error in Laplace approximation is controllable as long as the true model size is $o(n^{1/3})$.
- Simulations show method is robust in terms of the choice of q.

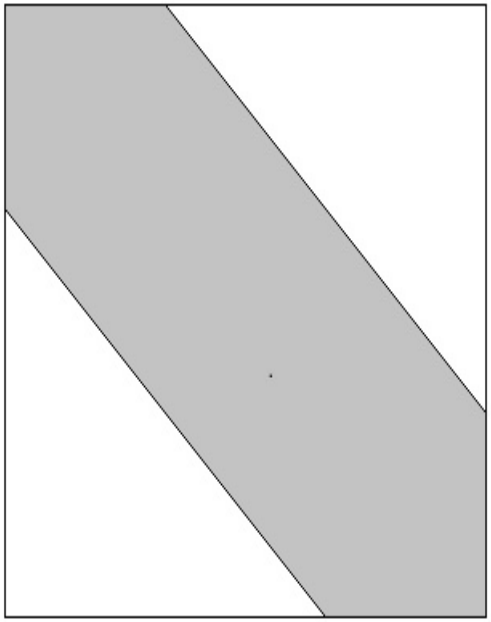
Nonparametric Additive Regression (contd.)

Table: Table corresponding to AR(1) predictors, $p = 500$, $r = 5$ and $q = 0.2$, choosing penalty parameter λ using penalized marginal likelihood criterion

	n	Error I	Error II	True.model
Approx.Bayes	100	2.335 (0.076)	0.120 (0.025)	0.040 (0.009)
Reich.method	100	0.980 (0.009)	392.020 (0.093)	0
G.Lasso	100	2.335 (0.076)	0.120 (0.025)	0.040 (0.009)
Approx.Bayes	200	1.460 (0.127)	0.060 (0.017)	0.110 (0.014)
Reich.method	200	1.330 (0.010)	356.610 (0.103)	0
G.Lasso	200	1.460 (0.127)	0.060 (0.017)	0.110 (0.014)
Approx.Bayes	500	0.405 (0.043)	0.175 (0.030)	0.540 (0.022)
Reich.method	500	-	-	-
G.Lasso	500	0.405 (0.043)	0.175 (0.030)	0.540 (0.022)

Estimation of Large Precision Matrix

- Consider multivariate Gaussian data $X \sim N_p(0, \Sigma)$.
- Let $\Omega = \Sigma^{-1}$ be the precision matrix.
- If the p variables are represented as the vertices of a graph G, then the absence of an edge between any two vertices j and j', which means conditional independence given others, is equivalent to $\omega_{jj'} = 0$.
- Roverato (2000), Letac and Massam (2007), Rajaratnam et al. (2008) studied families of conjugate priors for Ω for Gaussian graphical models called G-Wishart and W_{P_G}-Wishart families, and obtained expressions for posterior mean when "G is decomposable with a perfect ordering of cliques".
- Marginal likelihood can also be calculated explicitly giving posterior distribution of the banding parameter k.



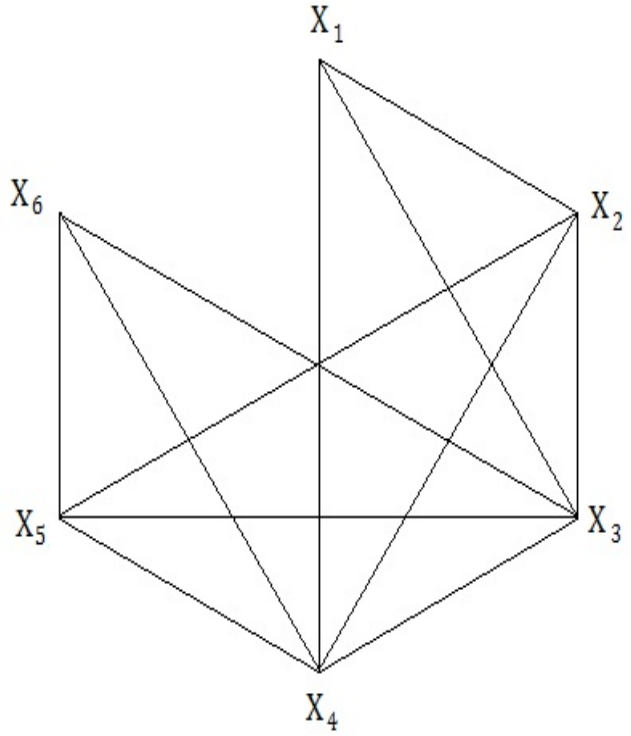


Figure: [Left] Structure of a banded precision matrix with shaded non-zero entries. [Right] The graphical model corresponding to a banded precision matrix of dimension 6 and banding parameter 3.

Posterior Convergence Rate for a Large Precision Matrix

- We study convergence rate under ℓ_∞-operator norm.
- We do not assume that the true Ω is a banded matrix, but only that it is approximable by banded matrices in the following sense: $\max_j \sum_i \{\omega_{jj'} : |j - j'| > k\} \leq \gamma(k)$ for all k, and $0 < \epsilon_0 \leq \min \mathrm{eig}_j \Omega \leq \max \mathrm{eig}_j \Omega \leq \epsilon_0^{-1} < \infty$.

Theorem

The posterior distribution of Ω converges at the rate $\epsilon_{n,k} = \max\left\{k^2(n^{-1} \log p)^{1/2}, \gamma(k)\right\}$.
In particular, the posterior distribution is consistent if $k \to \infty$ such that $k^4 n^{-1} \log p \to 0$.

Steps in the Proof of Posterior Convergence

- Conjugacy gives exact expressions for the posterior mean.
- We show that the posterior mean and the graphical MLE are k^2/n close.
- We show that the graphical MLE and the true Ω are $\epsilon_{n,k}$ close.
- We find posterior concentration around the posterior mean decomposing W_{P_G}-Wishart in Wishart over the cliques, representation of Wishart as sum of self-outer-product ZZ' of normal variables and applying exponential maximal inequalities.
- In all three steps, the most important issue is controlling the number of terms, which is of the order of p, but a closer look reveals that at any entry, at most $(2k+1)$ terms can be non-zero.

Graphical Lasso for Sparse Precision Matrix

- Developed in various papers — Meinshausen and Bühlman (2006), Yuan and Lin (2007), Banerjee et al. (2008), Friedman, Hastie and Tibshirani (2008).
- Minimize $\log\det\Omega - \mathrm{tr}(S\Omega) - \lambda\|\Omega\|_1$ subject to p.d. Ω, where S is the sample covariance matrix.
- Computation is doable in $O(p^3)$ steps by R package Glasso, but often has convergence issues. Uses a blockwise co-ordinate descent algorithm. For $10^3 \times 10^3$ matrix (approximately 500K parameters) takes 1 min.
- Improved algorithms by Mazumder and Hastie (2012), Witten et al. (2011).

Convergence rate of Graphical Lasso

- Convergence rate studied by Rothman et al. (2008). If $\lambda \asymp \sqrt{(\log p)/n}$, convergence rate in Frobenius (aka Euclidean) norm is $\sqrt{((p+s)\log p)/n}$, where s is the number of non-zero off-diagonal entries.
- Equivalently, in normalized Frobenius, the rate is $\sqrt{(\log p)/n}$ if $s = O(p)$. For the operator norm, the rate is $\sqrt{((s+1)\log p)/n}$.

Bayesian Graphical Lasso for Sparse Precision Matrix

- Wang (2012): Put independent exponential prior on diagonal entries, Laplace on off-diagonals, subject to positive definiteness restriction.
- Posterior mode is graphical Lasso.
- Not a real sparse prior. Posterior sits on non-sparse matrices, and hence cannot converge near the truth in high dimension.
- Real sparsity can be introduced by an extra Dirac mass at zero for off-diagonal entries.
- Computation becomes a challenge. Traditional MCMC/RJMCMC do not work in high dimensional setting.
- What can we say about posterior convergence rates?

Approximate Computation of Bayesian Graphical Lasso

- Use Laplace approximation around graphical Lasso restricted to the sparsity (as in the sparse linear/additive regression model), which is the posterior mode since the penalty is additive with minimum zero at zero.
- Explicit calculation of the Hessian possible.
- If $q < 1/2$, where q is the weight given to the non-singular part in the prior for the off-diagonal elements, then as before, non-regular models are dominated by the corresponding regular models in terms of posterior probability,
- Error in Laplace approximation can be controlled if $(p+s)\epsilon_n \to 0$, where ϵ_n is the posterior convergence rate in Frobenius norm.

Posterior Convergence Rate for Bayesian Graphical Lasso

- Use Frobenius norm $\|\Omega_0^{-1}\Omega - I\|_F$ on Ω scaled by the true Ω_0.
- This is comparable with the Hellinger distance between $N_p(0, \Omega_0)$ and $N_p(0, \Omega)$, the square root of their Kullback-Leibler divergence and the Euclidean norm on the vector of eigenvalues of $\Omega_0^{-1}\Omega$ centered by 1.
- Use general theory of posterior convergence rate [Ghosal, Ghosh and van der Vaart (2000)]. This needs bounding Hellinger entropy, assuring prior concentration in Kullback-Leibler sense and finally linking the Hellinger distance with the distance of interest.
- In view of the equivalence of distances, need bounding entropy and obtain prior concentration in terms of Frobenius norm.

Posterior Convergence Rate (contd.)

- Entropy calculation reduces to linking with Euclidean entropy.
- If eigenvalues of Ω_0 lie in $[a, b] \subset (0, \infty)$, calculation of prior concentration reduces to calculation of entry-wise prior concentration in view of a priori "independence" of the entries.
- Leads to the same convergence rate as the (non-Bayesian) graphical Lasso: $\epsilon_n = \sqrt{((p+s)\log p)/n}$.

Bayesian Density Regression

- First consider the situation of fixed dimension p and no sparsity.
- Model conditional density of Y given $X = x$ at y by a convex combination of tensor product of B-splines: $\sum \theta_{j,j_1,\ldots,j_p} B_j^*(y) B_{j_1}(x_1) \cdots B_{j_p}(x_p)$, $j, j_1, \ldots, j_p \leq J$, where B's are B-splines and B^*'s are normalized B-splines, where $\theta_{j,j_1,\ldots,j_p} \geq 0$ and add up to 1.
- Such functions approximate any C^α conditional density within $J^{-\alpha}$.
- A prior for f is obtained from the natural J^{p+1}-dimensional Dirichlet distribution on the θ-vector, and an exponential tailed infinitely supported prior on J.

Bayesian Density Regression: Computation

- Posterior mean can be analytically expressed as a sum of several explicit terms.
- This is because the likelihood in θ-vector is an explicit linear combination of a polynomial in θ-vector, and any polynomial in θ can be integrated out with respect to a Dirichlet distribution.
- Moreover the numerator and the denominator in the expression for posterior mean is similar looking, but the numerator contributes one extra factor in the likelihood.
- The number of such terms is very high.
- Sampling of terms is an option.

Convergence Rate for Bayesian Density Regression

- Using the general theory of posterior convergence [Ghosal, Ghosh and van der Vaart (2000)], in terms of averaged squared Hellinger distance on conditional densities (with respect to the distribution of the covariates), posterior converges at the optimal rate $n^{-\alpha/(2\alpha+p+1)}$ up to a log factor, for any (unknown) smoothness level, that is the posterior is automatically rate adaptive.
- Here every calculation reduces to Euclidean space.
- Entropy grows like $J^{2p+1}\log(1/\epsilon)$.
- Prior concentration is like $\epsilon^{-J^{p+1}}$.
- As long as $\epsilon \geq J^{-\alpha}$, the order of bias, the convergence rate is the solution of $n\epsilon^2 \asymp J^{p+1}\log(1/\epsilon)$, which is $n^{-\alpha/(2\alpha+p+1)}$ up to a log factor.

Bayesian Density Regression in High Dimension

- When p is high, but the conditional density only depends a fixed d co-ordinates only, the oracle rate is $n^{-\alpha/(2\alpha+d+1)}$.
- Introduce a variable selection step in the prior using an indicator γ_k for the inclusion of the kth variable.
- Further impose a bound on the total number of variables in the model that can grow only very slowly. like logarithmically, or impose a very strong tail condition on the prior for total model size. This ensures that high complexity models have very low prior probability.
- Still nearly analytic computation of posterior mean is possible.
- Appropriate modification of posterior convergence arguments leads to the adaptive oracle rate $n^{-\alpha/(2\alpha+d+1)}$ up to a logarithmic factor.

Bayesian Density Regression in High Dimension (contd.)

Table: Density regression example
$Y|X \sim \text{Beta}(5X_2 \exp(2X_1), 5X_3^2 + 3X_4)$.

Dim	$n = 100$				$n = 500$			
	rs (l_1)	rs (l_2)	ls (l_1)	ls (l_2)	rs (l_1)	rs (l_2)	ls (l_1)	ls (l_2)
5	.65	.61	.73	.85	.70	.67	.70	.77
10	.66	.59	.78	.92	.74	.76	.81	1.14
50	.67	.63	.65	.66	.74	.74	.73	.84
100	.70	.68	.70	.78	.66	.62	.65	.69
500	.58	.50	.69	.80	.77	.83	.78	1.16
1000	.74	.73	.75	1.14	.66	.61	.81	1.17
s.e.	.04	.08	.06	.12	.05	.09	.08	.18