Efficient statistical inference for high dimensional and nonparametric models

Natalia Bochkina

School of Mathematics and Maxwell Institute

N.Bochkina@ed.ac.uk

13 October 2014

High dimensional data and modern statistics

Availability of noisy high dimensional data from biology and medicine (genomics, genetics, tomography, brain imaging), ecology, social networks and other data (netflix problem) triggered development of **"high p small n"** paradigm in statistics where the number of unknowns p is higher than the sample size n.

Classical statistical inference works if p is fixed as $n \to \infty$.

Current data has $p \to \infty$ as $n \to \infty$, often $p/n \to const$ or $p/n \to \infty$.

To ensure consistent and efficient inference, this required development of novel statistical methods, often embedding **a priori information** elicited from experts.

- Main statistical methods: penalised likelihood and Bayesian models
- Computational challenges: to implement the methods efficiently for a large number of unknowns
- Mathematical challenges: to come up with statistical inference methods that guarantee consistency and efficiency

It is often of interest to recover structure in these types of data.

High dimensional data and modern statistics

Availability of noisy high dimensional data from biology and medicine (genomics, genetics, tomography, brain imaging), ecology, social networks and other data (netflix problem) triggered development of **"high p small n"** paradigm in statistics where the number of unknowns p is higher than the sample size n.

Classical statistical inference works if p is fixed as $n \to \infty$.

Current data has $p \to \infty$ as $n \to \infty$, often $p/n \to const$ or $p/n \to \infty$.

To ensure consistent and efficient inference, this required development of novel statistical methods, often embedding **a priori information** elicited from experts.

- Main statistical methods: penalised likelihood and Bayesian models
- Computational challenges: to implement the methods efficiently for a large number of unknowns
- Mathematical challenges: to come up with statistical inference methods that guarantee consistency and efficiency

It is often of interest to recover structure in these types of data.

High dimensional data and modern statistics

Availability of noisy high dimensional data from biology and medicine (genomics, genetics, tomography, brain imaging), ecology, social networks and other data (netflix problem) triggered development of **"high p small n"** paradigm in statistics where the number of unknowns p is higher than the sample size n.

Classical statistical inference works if p is fixed as $n \to \infty$.

Current data has $p \to \infty$ as $n \to \infty$, often $p/n \to const$ or $p/n \to \infty$.

To ensure consistent and efficient inference, this required development of novel statistical methods, often embedding **a priori information** elicited from experts.

- Main statistical methods: penalised likelihood and Bayesian models
- Computational challenges: to implement the methods efficiently for a large number of unknowns
- Mathematical challenges: to come up with statistical inference methods that guarantee consistency and efficiency

It is often of interest to recover structure in these types of data.

Statistical inference for high dimensional data

Likelihood: $\mathrm{Y} = (Y_1, \dots, Y_n) \sim f(\mathrm{Y} \mid \theta)$, for some $\theta \in \Theta \subseteq \mathbb{R}^p$, $p \gg n$.
Aim: to estimate unknown θ, its confidence region, make decisions.

Penalised log likelihood estimator:

$$\widehat{\theta} = \arg\min_{\widehat{\theta}} \left[-\log p(\mathrm{Y} \mid \widehat{\theta}) + pen(\widehat{\theta}) \right]$$

where the penalty reflects desirable properties of the solution, e.g. sparsity.
Problems:

- Construction of confidence regions for $\widehat{\theta}$ and other decision making.
- Assumptions are often not verifiable.

Bayesian model:
Given prior $p(\theta)$, posterior distribution is

$$\begin{aligned} p(\theta \mid \mathrm{Y}) &= \frac{f(\mathrm{Y} \mid \theta)\, p(\theta)}{\int_\Theta f(\mathrm{Y} \mid \theta)\, p(\theta)\, d\,\theta}, \\ \widehat{\theta} &= \arg\max_{\widehat{\theta}} \mathbb{E}\left(d(\widehat{\theta}, \theta) \mid \mathrm{Y} \right) \end{aligned}$$

given a distance d on Θ. Bayesian analogues of a confidence region and decision making can be constructed.

Challenges

- Computational: construct a fast algorithm to compute $\widehat{\theta}$ (and $p(\theta \mid \mathrm{Y})$).

- Mathematical: choose a penalty $pen(\theta)$ / prior $p(\theta)$ to ensure that $\widehat{\theta}$ is consistent and efficient, i.e.

$$\mathbb{E}[d(\widehat{\theta}, \theta)]^2 \to 0 \quad \text{as } n \to \infty,$$

  at the best possible rate.

For Bayesian inference, efficiency is related to local concentration of the posterior distribution around the true value of θ (Bernstein-von Mises theorem).

Research problems

- Modelling and decision making for genomic data, including model checks
  Bochkina et al (2006, 2007, 2010)
- Modelling dependence structure in genomics data and data integration
  A. Caballe, Bochkina, C.-D. Meyer
- Statistical inference for compound sparse high dimensional problems
  Bochkina & Ritov (2011)
- Concentration of posterior distribution (Bernstein – von Mises theorem) for nonregular and misspecified models, with application to tomography
  Bochkina and Green (2014), Bochkina (2013)
- Concentration of posterior distribution for semiparametric models with functional nuisance parameter
  Bochkina and Rousseau
- Adjusting Bayesian inference for approximated models

Concentration of posterior distribution (Bernstein–von Mises theorem)

For correctly specified regular models, as $n \to \infty$,

$$I_{\theta_{\text{true}}}^{1/2}(\theta - \theta_{\text{true}}) \mid \mathbf{Y} \sim N_p(\Delta_n, I_p)$$

where $I_{\theta_{\text{true}}}$ is Fisher information:

$$I_{\theta_{\text{true}}} = \mathbb{E}_{\theta_{\text{true}}} \left(\frac{\partial \log f(\mathbf{Y} \mid \theta_{\text{true}})}{\partial \theta} \right)^2.$$

That is, Bayesian inference in asymptotically optimal in frequentist sense.

For iid models, $I_{\theta_{\text{true}}} = n i_{\theta_{\text{true}}}$ leading to standard $\sqrt{n}$ parametric convergence rate.

For misspecified regular models, where $f_{true}(\mathbf{Y}) \notin \{f(\mathbf{Y} \mid \theta), \theta \in \Theta\}$,

$$V_{\text{true}}^{1/2}(\theta - \theta_{\text{true}}) \mid \mathbf{Y} \sim N_p(\Delta_n, I_p)$$

where

$$V_{\text{true}} = \mathbb{E}_{\text{true}} \left(\frac{\partial \log f(\mathbf{Y} \mid \theta^\star)}{\partial \theta} \right)^2,$$

and $\theta^\star$ corresponds to the model $f(\mathbf{Y} \mid \theta^\star)$ closest to the true one $f_{true}(\mathbf{Y})$ in Kullback-Leibler distance.

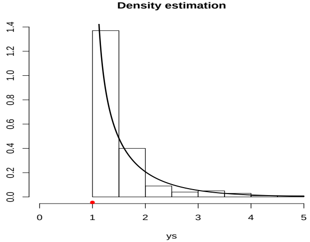
For nonregular models, where θ_{true} or $\theta^\star$ are on the boundary of Θ,

$$n(\theta - \theta^\star)_j \mid \mathbf{Y} \sim \Gamma(\alpha, b_j) \quad \text{as } n \to \infty$$

for some directions j where α is parameter of the prior distribution.

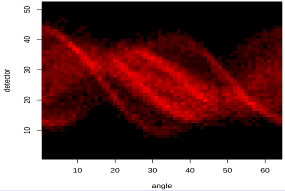
Application to nonparametric and inverse problems, tomography

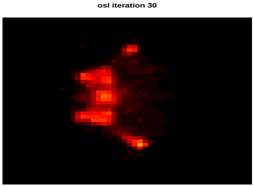
- Density estimation: $(Y_1, \ldots, Y_n)$ is a sample from density f.
- Nonparametric regression: $(Y_1, \ldots, Y_n)$ are noisy observations of f at points $(t_1, \ldots, t_n)$.
- Inverse problem: $(Y_1, \ldots, Y_n)$ are noisy observations of $\mathcal{A}(f)$ (indirect observations of f) at points $(t_1, \ldots, t_n)$.



Aim: estimate unknown function f.

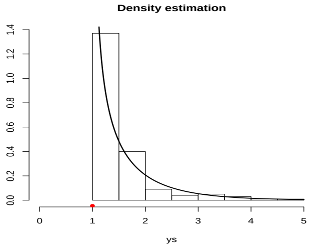
Inverse problem, tomography (plots by P.J.Green)





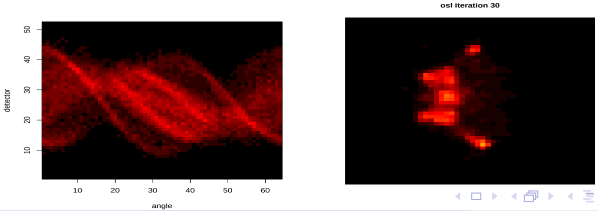
Application to nonparametric and inverse problems, tomography

- Density estimation: $(Y_1, \ldots, Y_n)$ is a sample from density f.
- Nonparametric regression: $(Y_1, \ldots, Y_n)$ are noisy observations of f at points $(t_1, \ldots, t_n)$.
- Inverse problem: $(Y_1, \ldots, Y_n)$ are noisy observations of $\mathcal{A}(f)$ (indirect observations of f) at points $(t_1, \ldots, t_n)$.



Aim: estimate unknown function f.

Inverse problem, tomography (plots by P.J.Green)



Adjusting Bayesian inference for approximated models

For complex Bayesian models, approximate models are often fitted to speed up the computation.

They often underestimate uncertainty in the posterior distribution.

The idea is

- to view approximate models as misspecified models
- to use BvM for misspecified models to adjust the inference

Adjusting Bayesian inference for approximated models

For complex Bayesian models, approximate models are often fitted to speed up the computation.

They often underestimate uncertainty in the posterior distribution.

The idea is

- to view approximate models as misspecified models
- to use BvM for misspecified models to adjust the inference