ROYAL INSTITUTE OF TECHNOLOGY

ListeningTalker

Regression Modeling and Estimation of Model Parameters

Petko N Petkov and W Bastiaan Kleijn

Sound and Image Processing Lab
KTH - Royal Institute of Technology

September 2, 2010

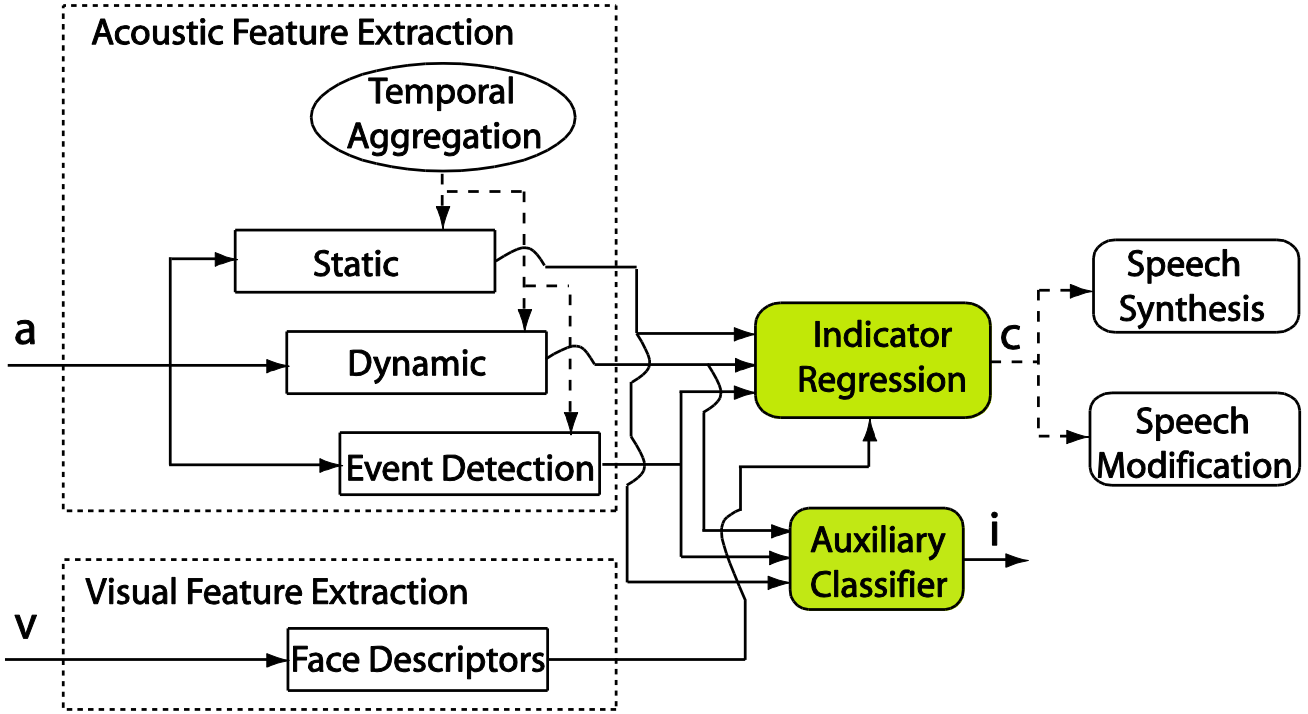
ROYAL INSTITUTE OF TECHNOLOGY

ListeningTalker

Outline

- Motivation
- The Regression Problem
- Model Parameter Estimation and Sparsity
  - Maximum Likelihood Approach
  - Bayesian Methods
    - Markov chain Monte Carlo approach
    - Variational approach
- Problem redefinition
- Summary

Motivation



LISTA system architecture

The regression model component - an integral part of the LISTA architecture!

ROYAL INSTITUTE OF TECHNOLOGY

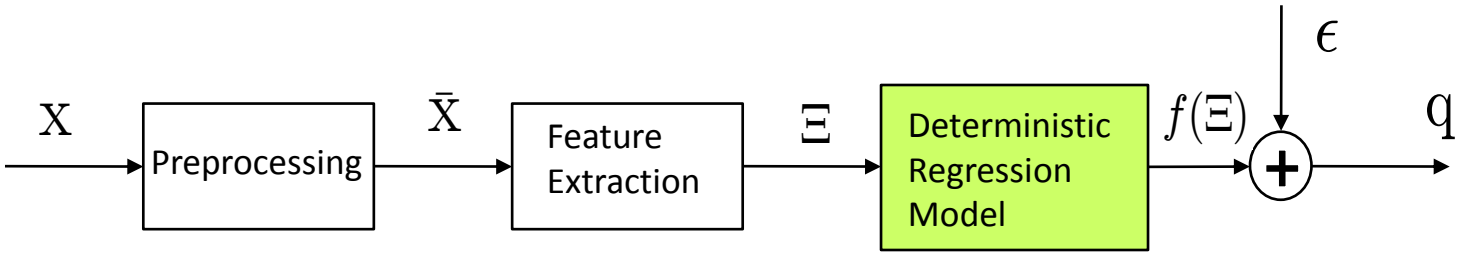
ListeningTalker

Progress

- Motivation
- The Regression Problem
- Model Parameter Estimation and Sparsity
  - Maximum Likelihood Approach
  - Bayesian Methods
    - Markov chain Monte Carlo approach
    - Variational approach
- Problem redefinition
- Summary

The Regression Problem

A regression constitutes a mapping from the continuous/discrete space of measurements to the continuous space of indicators.



- **Preprocessing** – calibrating effect on features
- **Feature extraction**
  - Sufficient statistic – discard irrelevant information
  - Facilitates parameter estimation - training data is limited
- **Regression model** – mapping

The Regression Problem

Formal definition from statistical perspective:

Additive model:

$$q_j = f(\Xi_j) + \epsilon_j,\ \epsilon \sim N(0, \sigma^2)$$

$$\hat{f}(\Xi_j) = \sum_{i=1}^{k} a_i B_i(\Xi_j)$$

Estimating the model parameters

- Frequentist (maximum likelihood) inference
- Bayesian inference

The Regression Problem

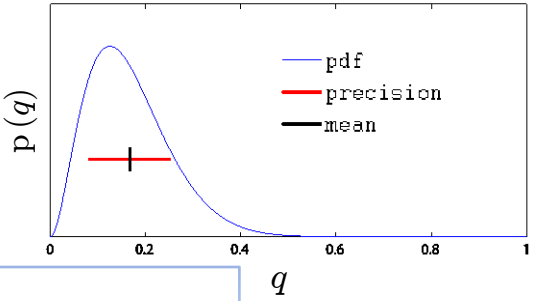
Direct model – an example:

$$\mathrm{p}\left(q_j|\mu_j,\,\phi_j\right)=\frac{1}{\mathrm{B}(\mu_j\phi_j,\,(1-\mu_j)\phi_j)}q_j^{\mu_j\phi_j-1}\left(1-q_j\right)^{(1-\mu_j)\phi_j-1}$$

$$\mu_j=\frac{e^{m(\Xi_j)+\epsilon_j}}{1+e^{m(\Xi_j)+\epsilon_j}},\quad \hat{m}\left(\Xi_j\right)=\mathrm{c}^{\mathrm{T}}\Xi_j,\quad \epsilon\sim N\left(0,\,\sigma^2\right)$$

$$\phi_j=e^{g(\Xi_j)+\xi_j},\quad \hat{g}\left(\Xi_j\right)=\mathrm{d}^{\mathrm{T}}\Xi_{\mathrm{j}},\quad \xi\sim N\left(0,\,\delta^2\right)$$

$\mu\in(0,\,1)\qquad \phi\in(0,\,\infty)$



Direct model:

- handles naturally data constraints, e.g., limited support
- takes into account that precision can vary

Progress

- Motivation
- The Regression Problem
- Model Parameter Estimation and Sparsity
  - Maximum Likelihood Approach
  - Bayesian Methods
    - Markov chain Monte Carlo approach
    - Variational approach
- Problem redefinition
- Summary

Model Parameter Estimaton

Assume limited support of the observations - reflects **LISTA**-related problems:

- **Quality and intelligibility assessment**
- **Mapping from environmental features to context indicators (pitch, energy, spectral tilt)**

Adopt the Beta regression model as a promising candidate

Consider the estimation of model parameters: $\{\mathrm{c},\, \mathrm{d},\, \sigma^2,\, \delta^2\}$

Consider the possibility that some of the coefficients are zero

- Improves prediction performance
- Facilitates interpretation

Investigate the advantages and disadvantages of various estimation techniques

Use quality assesment data for illustrative purposes

Data for solving LISTA-specifc regression problems comes from related WPs

ML Estimator

Assume independence between observations and no random effects

Problem becomes: $\operatorname{argmin}_{\mathrm{c},\,\mathrm{d}} - \sum_{i=1}^{N} \log\left(\mathrm{p}\left(q_i|\mathrm{c},\,\mathrm{d},\,\Xi_i\right)\right)$

Derivatives and Hessians of interest are readily obtained in closed-form

$$\frac{\partial \mathrm{LL}}{\partial c_j} = \sum_i \left\{ \left(\phi_i \left(\log\left(q_i\right) - \log\left(1-q_i\right)\right) - \psi_o\left(\mu_i\phi_i\right) + \psi_o\left(\phi_i - \mu_i\phi_i\right)\phi_i\right)\mu_i \left(1 + e^{\mathrm{c}^{\mathrm{T}}\mathrm{x}_i}\right)^{-1} x_{ij} \right\}$$

$$\frac{\partial \mathrm{LL}}{\partial d_j} = \sum_i \left\{ \left(\mu_i \log\left(q_i\right) + \left(1-\mu_i\right)\log\left(1-q_i\right) + \psi_o\left(\phi_i\right) - \psi_o\left(\mu_i\phi_i\right)\mu_i - \psi_o\left(\phi_i - \mu_i\phi_i\right)\left(1-\mu_i\right)\right)\phi_i x_{ij} \right\}$$

Optimization process is iterative due to coefficient interdependence

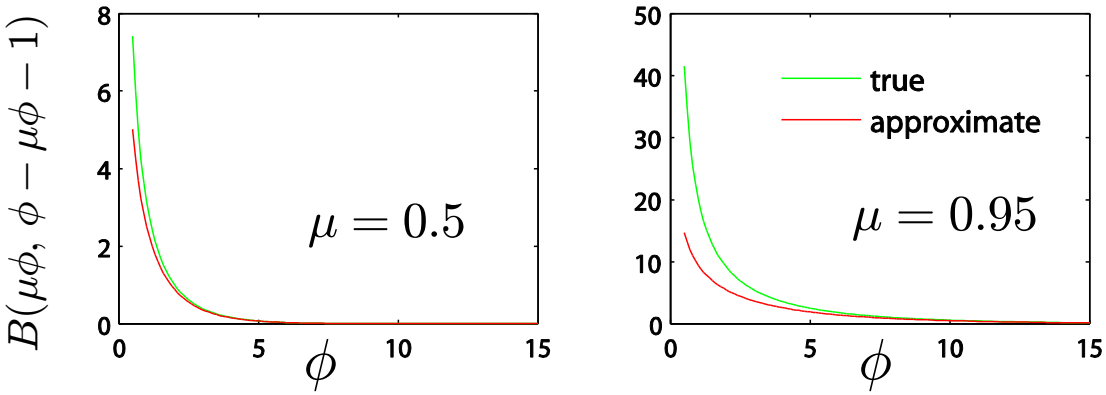
Suitable optimization algorithms are, e.g., quasi-Newton, interior points

Objective is **NOT convex** – initialization is very important

ML estimator - stability

Evaluating the Beta function and its logarithm can be unstable. Usage of Stirling's approximation addresses this problem:

$$B\left(\mu\phi,\,(1-\mu)\,\phi\right) \approx \sqrt{2\pi}\,\frac{(\mu\phi)^{\mu\phi-\frac{1}{2}}\left((1-\mu)\phi\right)^{(1-\mu)\phi-\frac{1}{2}}}{\phi^{\phi-\frac{1}{2}}}$$



Approximation accuracy:

- increases with the precision of the data
- higher in the middle of the support range

ML estimator - sparsity

Ideally **L0** norm should be used but this has ”combinatorial” complexity

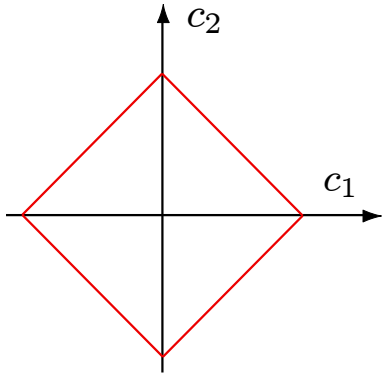
Sparsity in an optimization context - L1 norm constraint

$$\text{argmin}_{\mathbf{c},\,\mathbf{d}} - \sum_{i=1}^{N} \log\left[\text{p}\left(\text{q}_i|\mathbf{c},\,\mathbf{d},\,\Xi_i\right)\right]$$

$$\text{s.t.}\;\; B_c \geq \sum_{j=1}^{K_c} |c_j|\,,\;\; B_d \geq \sum_{j=1}^{K_d} |d_j|$$

Where optimal budget values B_c and B_d are determined by cross-validation

As budget decreases, coefficients go to zero



ML estimator - sparsity

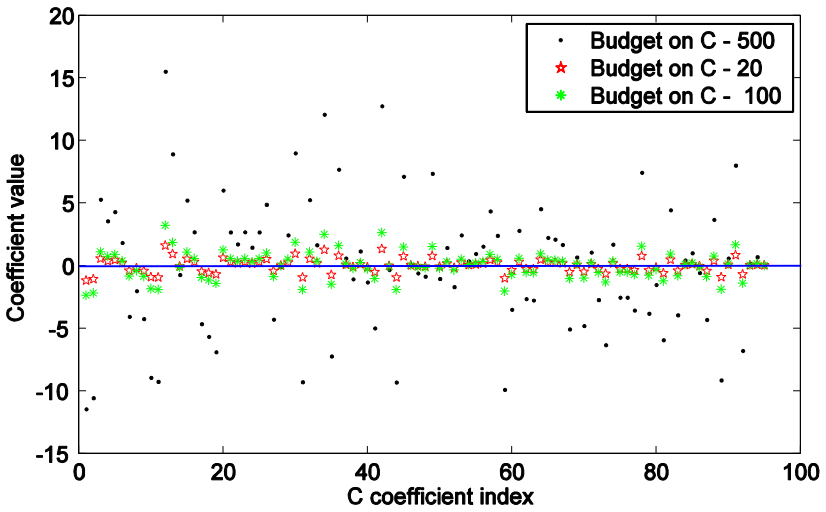
L1 norm is difficult to implement directly – relax problem

$$\text{argmin}_{\text{c}^+,\,\text{c}^-,\,\text{d}} - \sum_{i=1}^{N} \log\left(\text{p}\left(q_i|\text{c}^+,\,\text{c}^-,\,\text{d},\,\Xi_i\right)\right)$$

$$\text{s.t.} \quad \text{c}^+ \geq 0, \quad \text{c}^- \geq 0, \quad B_c \geq \sum_{j=1}^{K_c} \left(c_j^+ + c_j^-\right)$$

where $c_j^+ - c_j^- = c_j$

ML estimator - sparsity



L1 norm works well for convex objectives

Non-convex objectives and L1 norm constraints do not work together

- Changing the budget leads to moving between local optima
- Total least squares with L1-norm constraint not addressed in literature

L1 norm constraint for this model does not lead to efficient feature selection!

Bayesian estimator

In a Bayesian set-up all model parameters are considered random variables:

- Define a prior distribution over model parameters
  (random effects come in naturally through bias terms)

- Using Bayes rule, derive the posterior distribution

$$\mathrm{p}\left(\mathrm{c},\ \mathrm{d}|\mathrm{q}\right) = \frac{\mathrm{p}(\mathrm{q}|\mathrm{c},\mathrm{d})\mathrm{p}(\mathrm{c},\mathrm{d})}{\int\int \mathrm{p}(\mathrm{q}|\mathrm{c},\mathrm{d})\mathrm{p}(\mathrm{c},\mathrm{d})dcd\mathrm{d}}$$

- Use the posterior distribution in evaluating statistics of interest

Advantages:

- Using posterior distributions in regression avoids convexity issues
- Takes into account estimation uncertainty – improves regression performance

Disadvantages:

- Closed-form solutions obtainable in few cases only
- Requires approximations
  - sampling techniques – MCMC simulation
  - variational approximations

Bayesian estimator

We desire **conjugate** priors for **closed-from solution**

The Beta distribution does not have a known conjugate prior

Assume a priori Normality and independency:

$$\mathrm{p}\left(c_j|0,\,\iota_j^2\right)=\frac{1}{\sqrt{2\pi\iota_j^2}}e^{-\frac{1}{2\iota_j^2}c_j^2} \qquad \mathrm{p}\left(d_j|0,\,\cdot\,_j^2\right)=\frac{1}{\sqrt{2\pi\kappa_j^2}}e^{-\frac{1}{2\kappa_j^2}d_j^2}$$

Set prior variances sufficiently high for data-driven inference

Sparsity related issues:

- Bayesian methods avoid overfitting by taking into account estimation uncertainty
- Sparsity explicitly introduced by Laplace priors or suitable hyperpriors on a priori variances

Bayesian estimator- MCMC

Objective:

- sample from posterior distribution
- evaluate statistics of interest through stochastic integration

We set up a **Gibbs sampler** with two steps: one for set **c** and another for set **d**

Step 1: Sample from conditional posterior over **c**:

$$p(\mathbf{c}|\,\mathbf{d},\,\mathbf{q}) \propto \prod_i \left[\frac{1}{\mathrm{B}\left(\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{-\mathbf{c}^{\mathrm{T}}\Xi_i}},\,\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{\mathbf{c}^{\mathrm{T}}\Xi_i}}\right)} q_i^{\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{-\mathbf{c}^{\mathrm{T}}\Xi_i}}-1} (1-q_i)^{\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{\mathbf{c}^{\mathrm{T}}\Xi_i}}-1} \right] e^{-\frac{1}{2\iota^2}\mathbf{c}^{\mathbf{T}}\mathbf{I}\mathbf{c}}$$

Step 2: Sample from conditional posterior over **d**:

$$p(\mathbf{d}|\,\mathbf{c},\,\mathbf{q}) \propto \prod_i \left[\frac{1}{\mathrm{B}\left(\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{-\mathbf{c}^{\mathrm{T}}\Xi_i}},\,\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{\mathbf{c}^{\mathrm{T}}\Xi_i}}\right)} q_i^{\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{-\mathbf{c}^{\mathrm{T}}\Xi_i}}-1} (1-q_i)^{\frac{e^{\mathbf{d}^{\mathrm{T}}\Xi_i}}{1+e^{\mathbf{c}^{\mathrm{T}}\Xi_i}}-1} \right] e^{-\frac{1}{2\kappa^2}\mathbf{d}^{\mathbf{T}}\mathbf{I}\mathbf{d}}$$

Each step invokes a dedicated random-walk Metropolis-Hastings algorithm

After a burn-in period, samples represent dependent draws from desired posteriors

Bayesian estimator- MCMC

MCMC advantages:

- Can produce samples from any distribution
- Distribution need only be known up to a scale factor
- Asymptotically, the approximation converges to the true posterior

MCMC disadvantages:

- Convergence (end of burn-in) is difficult to determine
- Convergence can be slow for high-dimensional problems
- Dependence of consequtive draws imposes further computational load

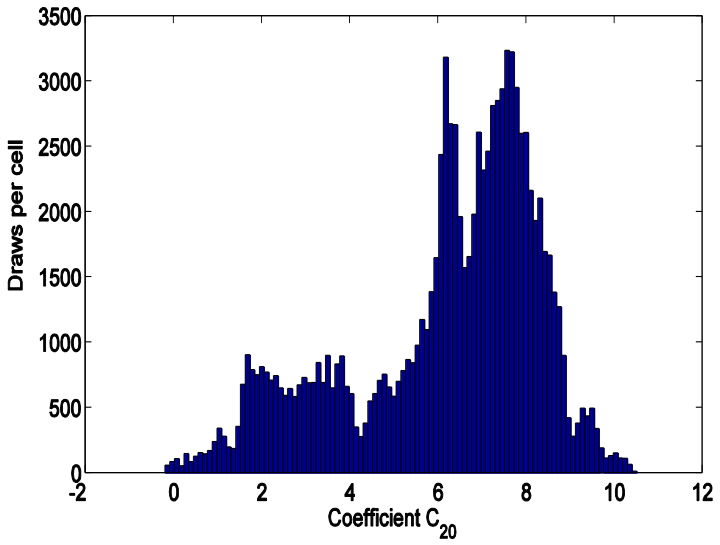
MH variations:

- Random walk MH – no prior info on the estimated parameter
  On acceptance of a sample, proposal distribution is centered at accepted value
  Method is **generally applicable** but incurs **high computational complexity**

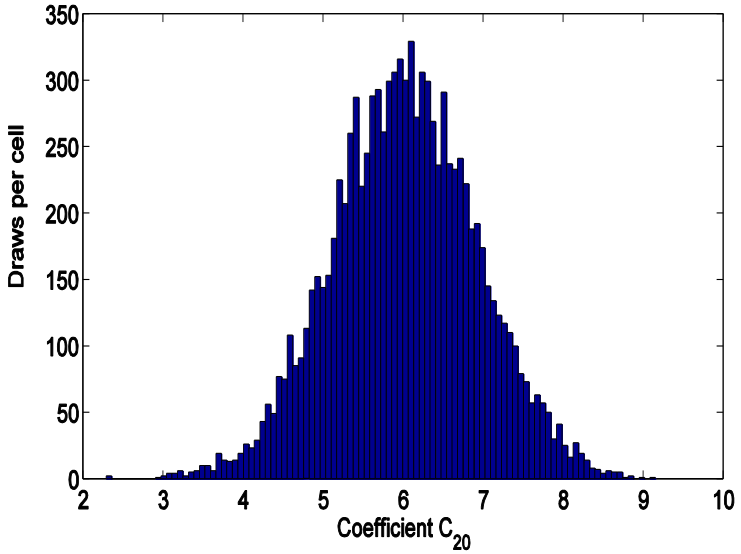
- Non-random walk – prior info on the estimated parameters is available
  Proposal distribution is centered
  Method is not generally applicable but more computationally efficient

Bayesian estimator- MCMC

Histograms for arbitrarily selected coefficient obtained with the two MH variations:



Random walk approach



Non-random walk approach

Note:

- much fewer samples are needed from the non-random walk approach
- posterior distribution resembles Gaussian – consider Laplace approximation or variatiotnal analysis

Performance Comparison

Correlation and root mean square for quality assessment data

database	r_{pc} ML	r_{pc} MCMC	ρ_{pc} ML	ρ_{pc} MCMC
BNR-X1	0.300	0.286	0.926	0.942
CNET-X1	0.318	0.340	0.912	0.875
NTT-X1	0.231	0.505	0.940	0.840
BNR-X3	0.266	0.300	0.944	0.900
CNET-X3	0.325	0.297	0.888	0.931
CSELT-X3	0.517	0.339	0.847	0.902
NTT-X3	0.312	0.239	0.887	0.936
Mean	**0.324**	**0.329**	**0.906**	**0.904**

Results indicate that initialization in ML solution was good

Variational Bayes estimator

Variational framework looks for a functional approximation on a constrained set

$$\log\left(\mathrm{p}\left(\mathrm{q}\right)\right) = \mathcal{L}\left(\mathrm{t}\right) + \mathrm{KL}\left(\mathrm{t}||\mathrm{p}\right)$$

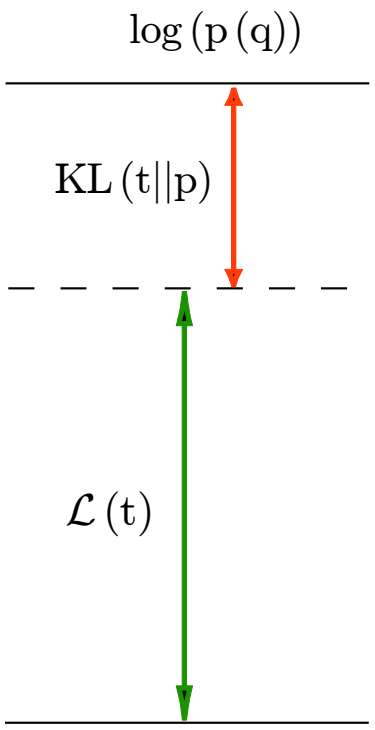
$$\mathcal{L}\left(\mathrm{t}\right) = \int \mathrm{t}\left(\theta\right) \log\left\{\frac{\mathrm{p}(\mathrm{q},\theta)}{\mathrm{t}(\theta)}\right\} \mathrm{d}\theta$$

$$\mathrm{KL}\left(\mathrm{t}||\mathrm{p}\right) = -\int \mathrm{t}\left(\theta\right) \log\left\{\frac{\mathrm{p}(\theta|\mathrm{q})}{\mathrm{t}(\theta)}\right\} \mathrm{d}\theta$$

θ can be any latent variable, $\theta = \{\mathrm{c},\ \mathrm{d}\}$

$\mathcal{L}\left(\mathrm{t}\right)$ - lower bound on the log marginal probability

$\mathrm{t}\left(\theta\right)$ - approximates the posterior of interest



Variational Bayes estimator

Different approximations are feasible, among them is the factorized posterior

$$\mathrm{t}(\theta) = \prod_i \mathrm{t}_i(\theta_i)$$

Substitute $\mathrm{t}(\theta)$ in $\mathcal{L}(\mathrm{t})$ and maximize w.r.t. t_j

Turns out, **bound is convex** w.r.t. the factors

- leads to an iterative process
- convergence is guaranteed due to convexity

At each iteration an **individual factor** is obtained as

$$\log\left(\mathrm{t}_j^*(\theta_j)\right) = E_{i\neq j}\left[\log\left(\mathrm{p}(\mathrm{q}, \theta)\right)\right] + \mathrm{const.}$$

Expectation is taken w.r.t. to the other factors that are kept fixed at current iteration

Approach is very attractive when resulting factors belong to known distributions

Variational Bayes estimator

The factorized variational approach is not generally suitable

Consider the Beta regression model

$$\mathrm{p}\left(\mathrm{q},\, \theta\right) = \mathrm{p}\left(\mathrm{q}|\mathrm{c},\, \mathrm{d}\right)\mathrm{p}\left(\mathrm{c}\right)\mathrm{p}\left(\mathrm{d}\right)$$

where similar to the MCMC case we adopt Normal priors

VB for coefficient c_j leads to

$$\log\left(\mathrm{t}_j^*\left(\theta_j \equiv c_j\right)\right) = E_{\neq c_j}\left\{-\sum\log\left[\mathrm{B}\left(\mu_i\phi_i,\, \left(1-\mu_i\right)\phi_i\right) + \sum_i \mu_i\phi_i\log\left(q_i\right)\right] + \sum_i\left(1-\mu_i\right)\phi_i\log\left(1-q_i\right)\right\} - \frac{c_j^2}{2\iota_j^2} + \mathrm{const.}$$

which is difficult to evaluate and does not lead to a known distribution

Local approximations are possible but very demanding and affect performance

ROYAL INSTITUTE OF TECHNOLOGY

listeningTalker

Outline

- Motivation
- The Regression Problem
- Model Parameter Estimation and Sparsity
  - Maximum Likelihood Approach
  - Bayesian Methods
    - Markov chain Monte Carlo approach
    - Variational approach
- Problem redefinition
- Summary

Problem redefinition

Estimation of the parameters in the Beta regression poses serious challenges

- **non-convexity for the ML approach**
- **no closed form solution and high computational load for the Bayesian approach**

As an alternative approach, consider modification of the original problem

A distribution resembling the Beta distribution is the Logit-Normal distribution

$$\mathrm{p}(q) = \frac{1}{\sqrt{2\pi}\sigma q(1-q)} e^{-\frac{1}{2}\left[\frac{\log\left(\frac{q}{1-q}\right)-\mu}{\sigma}\right]^2}, \quad q \in (0,\ 1)$$

The logistic trasnform $\tilde{q} = \log\left(\frac{q}{1-q}\right)$ leads to the Normally-distributed responses

$$\tilde{q} \sim N\left(\mu, \sigma^2\right)$$

Problem redefinition

Strategy: transform original observations using the logit transform

- **Solve regression problem in transform domain**
- **Transform back to original domain for performance evaluation**

Advantages:

- Advanced regression models for Normally-distributed variables are in large supply
- The likelihood function can be shown to be concave – global maximum can be found
- Bayesian analysis is facilitated as we can use conjugate priors (at least for **c**)

Disadvantages:

- Moments of the logistic-normal distribution are not trivial to obtain
- Sampling and stochastic integration are commonly used
- Optimality in one domain does not guarantee optimality in the other – sparsity through L1-norm constraint

Problem redefinition-ML

Convexity of the ML problem in transform domain – proof

$$\log(\tilde{\mathrm{q}}) = \log\left(\prod_i p\left(\tilde{q}_i|\mu_i,\,\sigma_i^2\right)\right) = \sum_i \log\left(p\left(\tilde{q}_i|\mu_i,\,\sigma_i^2\right)\right)$$

$$p\left(\tilde{q}_i|\mu_i,\,\sigma_i^2\right) = \frac{1}{(2\pi)^{\frac{1}{2}} e^{\mathrm{d}^{\mathrm{T}}\Xi_i}} e^{-\frac{1}{2e^{2\mathrm{d}^{\mathrm{T}}\Xi_i}}\left(\tilde{q}_i - \mathrm{c}^{\mathrm{T}}\Xi_i\right)^2}$$

$$\log(\tilde{\mathrm{q}}) = -0.5\sum_i \log(2\pi) - \sum_i \mathrm{d}^{\mathrm{T}}\Xi_i - 0.5\sum_i e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\left(\tilde{q}_i^2 - 2\tilde{q}_i\mathrm{c}^{\mathrm{T}}\Xi_i + \mathrm{c}^{\mathrm{T}}\Xi_i\Xi_i^{\mathrm{T}}\mathrm{c}\right)$$

Method:

- **Fix d and show convexity in c**
- **Fix c and show convexity in d**
- **Conclude convexity of the problem and existence of unique optimum**

Problem redefinition-MIL

Problem convexity in **c** (fixed **d**) – isolate all terms dependent on **c**

$$A \equiv -0.5 \sum_i e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i} \left(\mathrm{c}^{\mathrm{T}}\Xi_i\Xi_i^{\mathrm{T}}\mathrm{c} - 2\tilde{q}_i\mathrm{c}^{\mathrm{T}}\Xi_i\right)$$

$$\equiv \sum_i \left\{-0.5e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\mathrm{c}^{\mathrm{T}}\Xi_i\Xi_i^{\mathrm{T}}\mathrm{c} + e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\tilde{q}_i\mathrm{c}^{\mathrm{T}}\Xi_i\right\}$$

$$\equiv \sum_i \mathrm{c}^{\mathrm{T}} \left\{-0.5\left(e^{-\mathrm{d}^{\mathrm{T}}\Xi_i}\Xi_i\right)\left(e^{-\mathrm{d}^{\mathrm{T}}\Xi_i}\Xi_i\right)^{\mathrm{T}}\right\}\mathrm{c} + \sum_i \left\{e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\tilde{q}_i\mathrm{c}^{\mathrm{T}}\Xi_i\right\}$$

$$\left(e^{-\mathrm{d}^{\mathrm{T}}\Xi_i}\Xi_i\right)\left(e^{-\mathrm{d}^{\mathrm{T}}\Xi_i}\Xi_i\right)^{\mathrm{T}} \succeq 0$$

Term I: negative semidefinite quadratic form - concave

Term II: linear form – convex/concave

The expression is concave in **c**

Problem redefinition-ML

Problem convexity in **d** (fixed **c**) – isolate all terms dependent on **d**

$$B \equiv -\sum_i \mathrm{d}^{\mathrm{T}}\Xi_i - 0.5\sum_i e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\left(\tilde{q}_i - \mathrm{c}^{\mathrm{T}}\Xi_i\right)^2$$

Term I: linear form – convex/concave

Term II: concave

- Composition rule: c*onvex non-decreasing function of a convex (linear) argument is convex*

  $e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}$ - *convex in* **d**
- *Due to the negative sign, the second term is concave in* **d**

To summarize:

- The objective is maximization of the log-likelihood
- Log-likelihood is concave in **c** and concave in **d**
- **The problem is convex, suggesting that the global optimum is obtained**

Problem redefinition-Bayesian

Normal likelihood - attractive in a Bayesian estimation framework

- **Conjugate prior for c**
- **Easier analysis**

$$\mathrm{p}\left(\mathrm{c},\,\mathrm{d}|\tilde{\mathrm{q}}\right) \propto \mathrm{p}\left(\tilde{\mathrm{q}}|\mathrm{c},\,\mathrm{d}\right)\mathrm{p}\left(\mathrm{c},\,\mathrm{d}\right)$$

$$\mathrm{p}\left(\tilde{\mathrm{q}}|\mathrm{c},\,\mathrm{d}\right) = \prod_i \frac{1}{(2\pi)^{\frac{1}{2}} e^{\mathrm{d}^{\mathrm{T}}\Xi_i}} e^{-\frac{1}{2e^{2\mathrm{d}^{\mathrm{T}}\Xi_i}}\left(\tilde{q}_i - \mathrm{c}^{\mathrm{T}}\Xi_i\right)^2}$$

$$\propto e^{-\frac{1}{2}\mathrm{c}^{\mathrm{T}}\left\{\sum_i e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\Xi_i\Xi^{\mathrm{T}}{}_i\right\}\mathrm{c}+\mathrm{c}^{\mathrm{T}}\left\{\sum_i \tilde{q}_i e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}\Xi_i\right\}} e^{-\sum_i \mathrm{d}^{\mathrm{T}}\Xi_i} e^{-\frac{1}{2}\sum_i \tilde{q}_i e^{-2\mathrm{d}^{\mathrm{T}}\Xi_i}}$$

The following observations can be made:

- The quadratic form in the exponential suggests multivariate Normal prior for **c**
- Conjugate prior for d is not obtained readily
- If variance is not parametrized in the data, Inverse-Gamma is the conjugate prior

Summary

- Motivated use of a regression model (RM)
- Selected an RM with high relevance to LISTA
- Introduced different estimation paradigms
  - Frequentist
  - Bayesian
- Introduced approximation techniques
  - MCMC simulation
  - Variational analysis
- Compared performance of different techniques
- Identified bottlenecks in presented estimation approaches
- Proposed a redefinition of the problem