Sparse multinomial logistic regression via approximate message passing

Evan Byrne

The Ohio State University

Joint work with Prof. Phil Schniter

Supported in part by NSF grant CCF-1218754 and NSF grant CCF-1018368.

Seminar @ OSU Laboratory for Artificial Intelligence

September 25th, 2015

Outline

- Overview of linear classification
- Algorithm details
- Numerical results

Motivating example

Micro-array gene expression data

- Can we identify which genes are good predictors of certain diseases?
- Given samples containing:
  - class label y indicating the type of disease (e.g., cancer)
  - feature vector $\boldsymbol{x}$ containing gene expression values
- How do we cope with $P \approx 10^4$ genes but only $N \approx 10^2$ samples?

Linear classification and feature selection

Linear classification: learn a weight matrix $\widehat{\boldsymbol{W}} \in \mathbb{R}^{P \times C}$ from training data $\{y_n \in \{1, ..., C\}, \boldsymbol{x}_n \in \mathbb{R}^P\}_{n=1}^N$ such that

$$y_n \approx \arg\max_i [\widehat{\boldsymbol{W}}^\mathsf{T} \boldsymbol{x}_n]_i,$$

and classification of unknown $\boldsymbol{x}_0$ via

$$\widehat{y}_0 = \arg\max_i [\widehat{\boldsymbol{W}}^\mathsf{T} \boldsymbol{x}_0]_i$$

has minimal error rate.

- Accurate classification when $N \ll P$ if "true" $\boldsymbol{W}$ is sufficiently sparse.
- Feature selection from largest elements in $\widehat{\boldsymbol{W}}$.
- How to design $\widehat{\boldsymbol{W}}$?

Multinomial logistic regression

One well known approach to multiclass linear classification is multinomial logistic regression (MLR).

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \sum_{n=1}^{N} \log q(y_n \,|\, \boldsymbol{x}_n, \boldsymbol{W}) + G(\boldsymbol{W})$$

where

$$q(y \,|\, \boldsymbol{x}, \boldsymbol{W}) = \frac{\exp([\boldsymbol{W}^{\mathsf{T}}\boldsymbol{x}]_y)}{\sum_{c=1}^{C} \exp([\boldsymbol{W}^{\mathsf{T}}\boldsymbol{x}]_c)}$$

and $G(\boldsymbol{W})$ is some concave regularization term, e.g., $-\lambda\|\boldsymbol{W}\|_1$.

Bayesian approach to MLR

Actual objective: minimize test-label error-rate
$\Longrightarrow$ equivalent to finding test-label posterior $p(y_0 \,|\, \boldsymbol{y}; \boldsymbol{X})$.

Assume $\exists$ "true" $\boldsymbol{W}$ and $\boldsymbol{y}$ corresponding to $\boldsymbol{X}$ s.t.

$$\boldsymbol{w}_p \sim p(\boldsymbol{w}_p)$$
$$y_n \,|\, \boldsymbol{W}^\mathsf{T}\boldsymbol{x}_n \sim q(y_n \,|\, \boldsymbol{W}^\mathsf{T}\boldsymbol{x}_n).$$

We can write the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{W}$ as

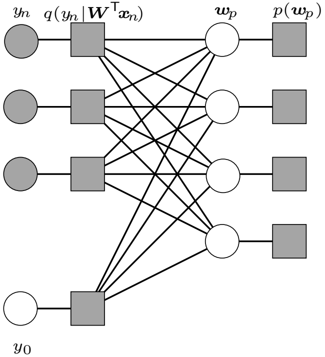
$$p(\boldsymbol{y}, \boldsymbol{W}; \boldsymbol{X}) \propto \prod_{n=1}^{N+T} q(y_n \,|\, \boldsymbol{W}^\mathsf{T}\boldsymbol{x}_n) \prod_{p=1}^{P} p(\boldsymbol{w}_p),$$

where $y_n, n \leq N$ are known, $y_n, n > N$ are unknown, and $\boldsymbol{w}_p$ is a row of $\boldsymbol{W}$.

Factor graph representation

$$p(\boldsymbol{y}, \boldsymbol{W}; \boldsymbol{X}) \propto \prod_{n=1}^{N+T} q(y_n \mid \boldsymbol{W}^\mathsf{T}\boldsymbol{x}_n) \prod_{p=1}^{P} p(\boldsymbol{w}_p)$$

$$\Updownarrow$$



- We could apply loopy belief propagation (based off the **Sum-Product** (SP) algorithm) to get approximate marginal posteriors (approximate due to loops in the factor graph), but...
- LBP is intractable due to the form of our distributions.
- Could try expectation-propagation, but infeasible due to large N and P.
- However, we can use previously developed approximate message passing (AMP) algorithms.

Approximate message passing

- AMP is derived from a simplification of message passing (sum-product or min-sum) that holds in the large system limit.
  - CLT to approximate messages as Gaussian.
  - Taylor series to reduce to $O(N+P)$ messages.

- The evolution of AMP:
  - AMP: for the linear model [Donoho, Maleki, Montanari 09].
  - Generalized-AMP (GAMP): for the generalized linear model with scalar variables [Rangan 11].
  - Hybrid-GAMP (HyGAMP): vector-valued extension of GAMP [Rangan, Fletcher, Goyal, Schniter 12].

- Since HyGAMP also approximates the **Min-Sum** (MS) algorithm, we can use it to solve the original MAP formulation of MLR.

Our contributions

1. Sparse multinomial logistic regression via HyGAMP.
   1. SP-HyGAMP: approximate minimum probability of error classifier.
   2. MS-HyGAMP: regularized MAP estimate $\widehat{\boldsymbol{W}}$ (focus on ℓ_1 case).
2. Simplified variants of both SP and MS-HyGAMP that are competitive with state-of-the-art algorithms w.r.t. algorithm runtime and test-error-rate.
3. Expectation-maximization (EM) and Stein's unbiased risk estimate (SURE) based methods to tune the model parameters online.

The HyGAMP algorithm for MLR

- Via approximate message passing, breaks one $P \times C$-dimensional inference problem into $N+P$ C-dimensional inference problems (but iterative).
- Messages take the form of C-dim normal distributions.
- Approximates the posterior of $\boldsymbol{w}_p$ and 'hidden' $\boldsymbol{z}_n \triangleq \boldsymbol{W}^{\mathsf{T}} \boldsymbol{x}_n$ as product of Gaussian and $p(\boldsymbol{w}_p)$ or $q(y_n \,|\, \boldsymbol{W}^{\mathsf{T}} \boldsymbol{x}_n)$, respectively.
- Each iteration is a series of linear steps and inference steps.
  - SP and MS-HyGAMP have identical linear steps.
  - Inference steps find mean/mode of approx. posteriors.

The HyGAMP algorithm for MLR

Require: Mode $\in$ {Sum-Prod, Min-Sum}, $\boldsymbol{y}$, $\boldsymbol{X}$, prior $p(\boldsymbol{w})$, inits. $\widehat{\boldsymbol{w}}_p$, $\boldsymbol{Q}_p^{\mathsf{w}}$.
Ensure: $t \leftarrow 0$; $\widehat{\boldsymbol{s}}_n(0) \leftarrow \boldsymbol{0}$.

repeat

- $\forall n : \boldsymbol{Q}_n^{\mathsf{p}} \leftarrow \sum_p X_{np}^2 \boldsymbol{Q}_p^{\mathsf{w}}(t)$
- $\forall n : \widehat{\boldsymbol{p}}_n \leftarrow \sum_p X_{np} \widehat{\boldsymbol{w}}_n - \boldsymbol{Q}_n^{\mathsf{p}} \widehat{\boldsymbol{s}}_n$
- **if** Min-Sum **then** (inference steps)
  - $\forall n : \widehat{\boldsymbol{z}}_n \leftarrow \arg\max_{\boldsymbol{z}} \log q(y_n \mid \boldsymbol{z}) \mathcal{N}(\boldsymbol{z}; \widehat{\boldsymbol{p}}_n, \boldsymbol{Q}_n^{\mathsf{p}})$
  - $\forall n : \boldsymbol{Q}_n^{\mathsf{z}} \leftarrow \left[-\frac{\partial^2}{\partial \boldsymbol{z}^2} \log q(y_n \mid \widehat{\boldsymbol{z}}_n) \mathcal{N}(\widehat{\boldsymbol{z}}_n; \widehat{\boldsymbol{p}}_n, \boldsymbol{Q}_n^{\mathsf{p}}) \right]^{-1}$
- **else if** Sum-Prod **then**
  - $\forall n : \widehat{\boldsymbol{z}}_n \leftarrow \mathrm{E}\left\{ q(y_n \mid \boldsymbol{z}) \mathcal{N}(\boldsymbol{z}; \widehat{\boldsymbol{p}}_n, \boldsymbol{Q}_n^{\mathsf{p}}) \right\}$
  - $\forall n : \boldsymbol{Q}_n^{\mathsf{z}} \leftarrow \mathrm{Cov}\left\{ q(y_n \mid \boldsymbol{z}) \mathcal{N}(\boldsymbol{z}; \widehat{\boldsymbol{p}}_n, \boldsymbol{Q}_n^{\mathsf{p}}) \right\}$
- **end if**
- $\forall n : \boldsymbol{Q}_n^{\mathsf{s}} \leftarrow [\boldsymbol{Q}_n^{\mathsf{p}}]^{-1} - [\boldsymbol{Q}_n^{\mathsf{p}}]^{-1} [\boldsymbol{Q}_n^{\mathsf{z}}] [\boldsymbol{Q}_n^{\mathsf{p}}]^{-1}$
- $\forall n : \widehat{\boldsymbol{s}}_n \leftarrow [\boldsymbol{Q}_n^{\mathsf{p}}]^{-1} (\widehat{\boldsymbol{z}}_n - \widehat{\boldsymbol{p}}_n)$
- $\forall p : \boldsymbol{Q}_p^{\mathsf{r}} \leftarrow \left[\sum_n X_{np}^2 \boldsymbol{Q}_n^{\mathsf{s}} \right]^{-1}$
- $\forall p : \widehat{\boldsymbol{r}}_p \leftarrow \widehat{\boldsymbol{w}}_p + \boldsymbol{Q}_p^{\mathsf{r}} \sum_n X_{np} \widehat{\boldsymbol{s}}_n$
- **if** Min-Sum **then** (inference steps)
  - $\forall p : \widehat{\boldsymbol{w}}_p \leftarrow \arg\max_{\boldsymbol{w}} \log p(\boldsymbol{w}) \mathcal{N}(\boldsymbol{w}; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}_p^{\mathsf{r}})$
  - $\forall p : \boldsymbol{Q}_p^{\mathsf{w}} \leftarrow \left[-\frac{\partial^2}{\partial \boldsymbol{w}^2} \log p(\widehat{\boldsymbol{w}}_p) \mathcal{N}(\widehat{\boldsymbol{w}}_p; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}_p^{\mathsf{r}}) \right]^{-1}$
- **else if** Sum-Prod **then**
  - $\forall p : \widehat{\boldsymbol{w}}_p \leftarrow \mathrm{E}\left\{ p(\boldsymbol{w}) \mathcal{N}(\boldsymbol{w}; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}_p^{\mathsf{r}}) \right\}$
  - $\forall p : \boldsymbol{Q}_p^{\mathsf{w}} \leftarrow \mathrm{Cov}\left\{ p(\boldsymbol{w}) \mathcal{N}(\boldsymbol{w}; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}_p^{\mathsf{r}}) \right\}$
- **end if**

until Terminated

Simplified HyGAMP (SHyGAMP)

- HyGAMP for MLR is computationally costly.
- We simplify by constraining message covariance matrices to be diagonal.
- Leads to:
  - faster matrix inversions
  - greatly simplifies inference steps
  - enables use of existing GAMPmatlab software

Online parameter tuning

- Model parameters require tuning.
  - SP: Bernoulli-Gaussian prior parameters $\{\alpha, \mu, \sigma^2\}$.
  - MS: ℓ_1 regularization parameter λ.
- We tune the SP parameters using EM [Vila, Schniter 13].
- We tune the MS parameter using a method based on Stein's unbiased risk estimate (SURE) [Mousavi, Maleki, Baraniuk 13].

SURE method to tune λ

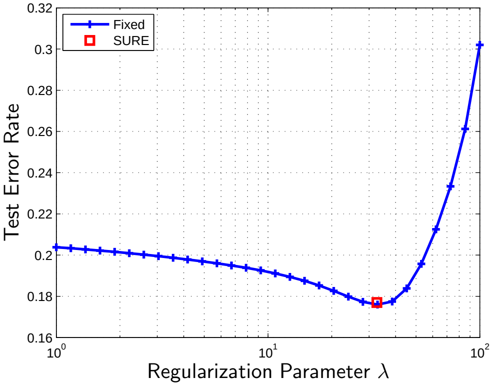
Basic idea: select λ to min Stein's unbiased risk estimate of the MSE.

- Recall: $\widehat{\boldsymbol{w}}_p = \arg\max_{\boldsymbol{w}} \log \mathcal{N}(\boldsymbol{w}; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}_p^{\mathsf{r}}) p(\boldsymbol{w}; \lambda)$.
- Assume $\widehat{\boldsymbol{r}} = \boldsymbol{w} + \sigma\, \boldsymbol{v}$ where $\widehat{\boldsymbol{r}}$ and σ are known.
  - Then, $\mathrm{E}\{S(\widehat{\boldsymbol{r}}; \sigma, \lambda)\} = \mathrm{E}\{|\widehat{\boldsymbol{w}} - \boldsymbol{w}|^2\}$.
  - Choose $\widehat{\lambda} = \arg\min_{\lambda} S(\widehat{\boldsymbol{r}}; \sigma, \lambda)$.
- However... the objective $S(\cdot)$ is non-smooth and has many local minima.
  - Prior work proposed approximate gradient descent, but too slow.
- We replaced an empirical average with a statistical average.
  - Smooth, easy to compute gradient.
  - Can now efficiently minimize via bisection.

Experimental validation of SURE

Synthetic data

- $C = 4,\ N = 300,\ P = 30\,000$
- $\boldsymbol{x}_n \mid (y_n = c) \sim \mathcal{N}(\boldsymbol{\mu}_c, v\boldsymbol{I})$
- Orthonormal $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$
- $K = 25$ discriminatory features
- Ran MS-SHyGAMP with fixed λ, then with SURE



Numerical Results

- Metrics
  - Classification accuracy
  - Algorithm runtime
- Target regime
  - High dimensional, and data-starved, i.e., $N \ll P$
  - Multiclass, $C \approx 10$
- Applications
  - Microarray gene-expression analysis
  - Text classification
  - Handwritten digit classification
- Competing algorithms
  - SBMLR [Cawley, Talbot, Girolami 07]
  - GLMNET [Friedman, Hastie, Tibshirani 10]

Gene-expression data

Sun *et al*

- Classes represent different types of glioma.
- $N = 179,\ P = 54\,613,\ C = 4$

Algorithm	% Error (SD)	Runtime (s)	$\widehat{K}_{99}$	$\|\widehat{\boldsymbol{W}}\|_0$
SP-SHyGAMP	32.0 (14.8)	**7.68**	10.29	218 452
MS-SHyGAMP	**30.9** (16.5)	12.33	31.04	49.25
SBMLR	32.3 (16.6)	24.10	48.41	72.41
GLMNET	31.1 (15.9)	32.30	24.79	39.28

Bhattacharjee *et al*

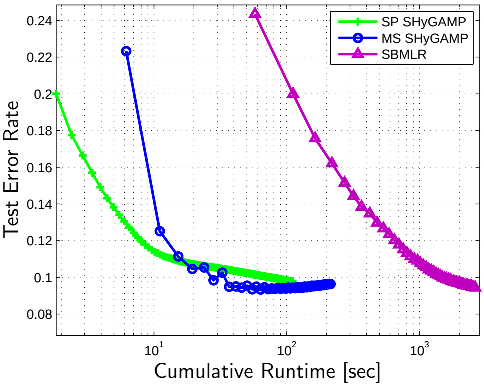
- Classes represent different types of lung carcinoma.
- $N = 203,\ P = 12\,600,\ C = 5$

Algorithm	% Error (SD)	Runtime (s)	$\widehat{K}_{99}$	$\|\widehat{\boldsymbol{W}}\|_0$
SP-SHyGAMP	8.0 (8.0)	**3.50**	14.64	63 000
MS-SHyGAMP	**6.2** (8.1)	8.04	40.62	66.29
SBMLR	6.6 (8.1)	7.36	46.55	79.68
GLMNET	6.6 (8.1)	13.96	53.17	93.50

Text classification

Reuter's Corpus Volume 1 (RCV1)

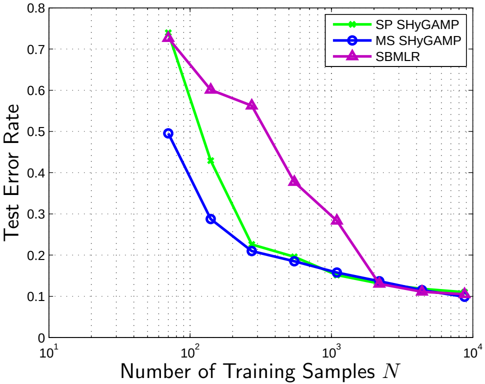
- Classes are the article's topic
- Features are frequency of keywords
- Sparse and non-zero mean $\boldsymbol{X}$
- $N = 14\,147$, $P = 47\,236$, $C = 25$
- Tested on $469\,571$ samples
- Plot test-error-rate vs algorithm runtime



Handwritten digit recognition

Mixed National Institute of Standards and Technology (MNIST)

- Classes are the digits 0-9
- Features pixels of an image ($P = 784$)
- We had in total $N = 70\,000$ samples
- Varied N from 70 to 10000



Summary

- Motivated by multiclass problems where $N \ll P$, want feature selection.
- Novel approach to sparse MLR by using message passing to break high dimensional inference problem into many smaller inference problems.
    - SP yields approximate marginal test label posteriors.
    - MS solves traditional ℓ_1-regularized objective.
- Automatically tune model parameters using EM and SURE techniques.
- Numerical results show we are competitive/superior to state-of-the-art.

References

- Paper: E. Byrne and P. Schniter, "Sparse multinomial logistic regression via approximate message passing," *arXiv:1509.04491*, Sep. 2015.
- GAMP: S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Thy.*, pp. 2168-2172, Aug. 2011.
- HyGAMP: S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter, "Hybrid generalized approximate message passing with applications to structured sparsity," in *Proc. IEEE Int. Symp. Inform. Thy.*, pp. 1236-1240, July 2012.
- EM: J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, pp. 4658-4672, Oct. 2013.
- SURE: A. Mousavi, A. Maleki, and R. G. Baraniuk, "Parameterless, optimal approximate message passing," *arXiv:1311:0035*, Nov. 2013.

Thank you

Questions?

SP-HyGAMP for MLR

Marginal posteriors are approximated as

"weights": $p(\boldsymbol{w}_p \,|\, \widehat{\boldsymbol{r}}_p; \boldsymbol{Q}^{\mathsf{r}}_p) \propto p(\boldsymbol{w}_p)\mathcal{N}(\boldsymbol{w}_p; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}^{\mathsf{r}}_p)$

"scores": $p(\boldsymbol{z}_n \,|\, y_n, \widehat{\boldsymbol{p}}_n; \boldsymbol{Q}^{\mathsf{p}}_n) \propto q(y_n \,|\, \boldsymbol{z}_n)\mathcal{N}(\boldsymbol{z}_n; \widehat{\boldsymbol{p}}_n, \boldsymbol{Q}^{\mathsf{p}}_n)$ for $\boldsymbol{z}_n \triangleq \boldsymbol{W}^{\mathsf{T}}\boldsymbol{x}_n$

Inference of weight vector $\widehat{\boldsymbol{w}}_p$

- Sparsity-promoting prior: $p(\boldsymbol{w}_p) = \alpha_0\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + (1-\alpha_0)\delta(\boldsymbol{w}_p)$
- Must compute $\widehat{\boldsymbol{w}}_p = \mathrm{E}\{p(\boldsymbol{w}_p \,|\, \widehat{\boldsymbol{r}}_p; \boldsymbol{Q}^{\mathsf{r}}_p)\}$, also covariance $\boldsymbol{Q}^{\mathsf{w}}_p$

Inference of score $\widehat{\boldsymbol{z}}_n$

- Must compute $\widehat{\boldsymbol{z}}_n = \mathrm{E}\{p(\boldsymbol{z}_n \,|\, y_n, \widehat{\boldsymbol{p}}_n; \boldsymbol{Q}^{\mathsf{p}}_n)\}$, also covariance $\boldsymbol{Q}^{\mathsf{z}}_n$.
  - Intractable due to form of $P(y_n \,|\, \boldsymbol{z}_n)$ (recall multinomial logistic function)
  - Solve via numerical integration (slow) or importance sampling (inaccurate)

MS-HyGAMP for MLR

Inference of weight vector $\widehat{\boldsymbol{w}}_p$

- Compute $\widehat{\boldsymbol{w}}_p = \arg\max_{\boldsymbol{w}} \log p(\boldsymbol{w}_p) + \log \mathcal{N}(\boldsymbol{w}_p; \widehat{\boldsymbol{r}}_p, \boldsymbol{Q}_p^{\mathsf{r}})$
- Under ℓ_1 regularization, i.e., Laplacian $p(\boldsymbol{w}_p)$

$$\widehat{\boldsymbol{w}}_p = \arg\max_{\boldsymbol{w}} -\frac{1}{2}(\boldsymbol{w} - \widehat{\boldsymbol{r}}_p)^{\mathsf{T}}[\boldsymbol{Q}_p^{\mathsf{r}}]^{-1}(\boldsymbol{w} - \widehat{\boldsymbol{r}}_p) - \lambda\|\boldsymbol{w}\|_1$$

- No CF solution, but can be solved iteratively using, e.g., minorization-maximization

Inference of score $\widehat{\boldsymbol{z}}_n$

$$\widehat{\boldsymbol{z}}_n = \arg\max_{\boldsymbol{z}} \log q(y_n \,|\, \boldsymbol{z}) - \frac{1}{2}(\boldsymbol{z} - \widehat{\boldsymbol{p}}_n)^{\mathsf{T}}[\boldsymbol{Q}_n^{\mathsf{p}}]^{-1}(\boldsymbol{z} - \widehat{\boldsymbol{p}}_n)$$

- Convex, solved via Newton's method

GM approximation details

We expand on

$$\frac{1}{1+\exp(-z)} \approx \sum_{l=1}^{L} \Phi\Big(\frac{z-\mu_l}{\sigma_l}\Big).$$

Change of variables:

$$q(y \,|\, \boldsymbol{z}) = \frac{1}{1+\sum_{c\neq y}\exp(-\gamma_c^y)},\ \gamma_c^y = z_y - z_c.$$

Gaussian mixture approximation:

$$\frac{1}{1+\sum_{c\neq y}\exp(-\gamma_c^y)} \approx \sum_{l=1}^{L} \alpha_l \prod_{c\neq y} \Phi\Big(\frac{\gamma_c-\mu_{cl}}{\sigma_{cl}}\Big).$$

SURE details

Input soft thresholding:

$$\widehat{w}_{pc} = f(\widehat{r}_{pc}, q^{\mathsf{r}}; \lambda) = \mathsf{sign}(\widehat{r}_{pc}) \max\{0, |\widehat{r}_{pc}| - \lambda q^{\mathsf{r}}\}.$$

Shifted estimation function

$$g(\widehat{r}, q^{\mathsf{r}}; \lambda) = f(\widehat{r}, q^{\mathsf{r}}; \lambda) - r.$$

Stein's result

$$\mathrm{E}\{|\widehat{w} - \mathsf{w}|^2\} = q^{\mathsf{r}} + \mathrm{E}\{g^2(r, q^{\mathsf{r}}; \lambda) + 2q^{\mathsf{r}} g'(r, q^{\mathsf{r}}; \lambda)\}.$$

We know $\mathrm{E}\{\mathsf{S}(\mathsf{r}, q^{\mathsf{r}}; \lambda\} = \mathsf{MSE}(\lambda)$, so we choose λ to minimize $\mathrm{E}\{S(\mathsf{r}, q^{\mathsf{r}}; \lambda)\}$.

SURE details con't

Minimize empirical average

$$\widehat{\lambda} = \arg\min_{\lambda} \sum_{p=1}^{P} \sum_{c=1}^{C} g^2(\widehat{r}_{pc}, q^{\mathsf{r}}; \lambda) + 2q^{\mathsf{r}} g'(\widehat{r}_{pc}, q^{\mathsf{r}}; \lambda).$$

Above is difficult, so instead solve

$$\widehat{\lambda} = \arg\min_{\lambda} \mathrm{E}\{g^2(\mathsf{r}, q^{\mathsf{r}}; \lambda) + 2q^{\mathsf{r}} g'(\mathsf{r}, q^{\mathsf{r}}; \lambda)\},$$

where $p(r) = \sum_{l=1}^{L} \alpha_l \Phi\left(\frac{r-\mu_l}{\sigma_l}\right)$.