Constrained Relative Entropy Minimization with Applications to Multitask Learning

Dissertation Defense

Oluwasanmi Koyejo

PhD Candidate
Department of Electrical and Computer Engineering
The University of Texas at Austin

April 29, 2013

Probabilistic Inference

Quantify uncertainty about x as $P(x)$.

Principle of maximum entropy

Choose $P(x)$ that captures known information and is as difficult as possible to discriminate from uniform[a] (Jaynes, 1957).

[a] Uniform distribution represents complete ignorance.

Principle of minimum discrimination information

Given $P(x)$, choose $Q(x)$ that captures new information and is as difficult to discriminate from $P(x)$ as possible (Kullback, 1959).

Comparison with Bayesian Inference

Why do we need (yet) another inference rule?

Bayesian Inference:

- Solid conceptual framework derived from axioms of probability.
- Provides rules for updating $P(x)$ given data samples $\mathcal{D}$.
- Can impose support constraints by defining appropriate $P(x)$.
- Support constraints are preserved by $P(x) \mapsto P(x|\mathcal{D})$.

But:

- Other kinds of structure (e.g. moment constraints) generally do not propagate.
- Bayesian inference does not provide a formal approach for updating P with new *constraint* information.

Prior Work

Relative entropy inference

- MAXENT for natural language processing (Berger et al., 1996).
- l_1 and l_2 norm ball constraints (Dudík et al., 2007).
- Banach norm constraints (Altun and Smola, 2006).
- Margin constraints ("probabilistic support vector machines"):
  - Maximum entropy discrimination (Jaakkola et al., 1999).
  - MED-LDA (Zhu et al., 2009).
  - Link prediction (Zhu, 2012).

Multitask Learning

An overview

Given multiple (related) learning tasks:

$$f_1 : x \mapsto y_1, \quad f_2 : x \mapsto y_2, \quad f_3 : x \mapsto y_3, \quad \ldots$$

Single task learning:

- Learn $f_1, f_2, f_3, \ldots$ separately.

Multiple task learning:

- Learn $F : x \mapsto Y$ jointly.
- Exploits inter-task relationships to improve performance
- When F is a bilinear function, MTL can be posed as a *matrix* estimation task.

Table of Contents

Main Contributions

Constrained Inference

A framework for probabilistic inference subject to expectation constraints:

- Constraints may be any combination of domain knowledge and data.
- Constraint sets may be nonconvex.
- Generalization of Bayesian inference that can incorporate additional structural constraints.

Main Contributions

Applications

- Constrained multitask learning and parameter estimation:
  - Applied to recommender systems and fMRI.
  - Rank constraints on the postdata mean matrix.
  - Kronecker constraints on the postdata covariance matrix.
  - Sparsity constraints on prior precision matrix.
- Nonparametric multitask learning:
  - Applied to predictive modeling of transposable data.
  - Nuclear norm constraints on the mean function.
- Multitask bipartite ranking:
  - Applied to disease-gene prioritization.
  - Ordering constraints on ranking variables.
  - Nuclear norm constraints on mean function of regression variables.

Table of Contents

Gaussian Process (GP)

- Collection of random variables $\{X(i),\ i \in \mathbb{I}\}$ completely described by:

$$\mu(i) = \mathrm{E}\left[\, X(i) \,\right],$$
$$\mathcal{C}(i,i') = \mathrm{E}\left[\, (X(i) - \mu(i))\big(X(i') - \mu(i')\big) \,\right].$$

- The collection $\mathbf{x} = \{x(i),\ \forall i \in \mathsf{J}\}$ is jointly Gaussian for any finite index set $\mathsf{J} \subset \mathbb{I}$.

- Matrix-variate Gaussian process (MV-GP) $\mathcal{MGP}\left(\phi, \mathcal{C}_{\mathsf{N}}, \mathcal{C}_{\mathsf{M}}\right)$ is a special case:
    - Double index $(m,n) \in \mathbb{M} \times \mathbb{N} = \mathbb{J}$.

$$\mathcal{C}((m,n),(m',n')) = \mathcal{C}_{\mathsf{M}}(m,m')\mathcal{C}_{\mathsf{N}}(n,n').$$

Relative Entropy (Kullback-Leibler divergence)

For probability measures $P, \ Q$

Let P be absolutely continuous with respect to Q, there exist probability density functions $p, \ q$ so that:

$$\mathrm{KL}(p\|q) = \int_{\mathbb{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

Some useful optimization-theoretic properties:

- Strictly convex wrt p.
- $\mathrm{KL}(p\|q) \geq 0$.
- $\mathrm{KL}(p\|q) = 0$ iff. $p = q$.

Exponential families

$$p_{\boldsymbol{\theta}}(x) = h(x)e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{t}(x)\rangle - A(\boldsymbol{\theta})}$$

- $\boldsymbol{\eta}(\boldsymbol{\theta})$ natural parameter
- $\mathbf{t}(x)$ sufficient statistics
- $h(x)$ base measure
- $A(\boldsymbol{\theta})$ log-partition function

The domain of the parameter $\boldsymbol{\theta}$ is a convex set defined as:

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} \;\middle|\; \int_{\mathbb{X}} h(x)e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{t}(x)\rangle} dx < \infty \right\}.$$

Constraints

We study constraints of the form $\mathrm{E}\left[\,\boldsymbol{\beta}(x)\,\right] \in \mathsf{C}$.

- $\boldsymbol{\beta}$ is called the *feature function*.
- C is a constraint set.

Examples:

- Empirical average (Jaynes, 1957): $\mathrm{E}\left[\,\boldsymbol{\beta}(x)\,\right] = \frac{1}{N}\sum_n \boldsymbol{\beta}(\hat{x}_n)$.
- Structural constraints (this dissertation):

$$\mathrm{E}\left[\,\boldsymbol{\beta}(x)\,\right] \geq 0, \quad \left\|\mathrm{E}\left[\,\boldsymbol{\beta}(x)\,\right]\right\|_0 \leq \epsilon.$$

Constrained Relative Entropy Minimization

Problem Statement

$$\inf_{q \in \mathcal{P}} \mathrm{KL}(q \| p) \text{ s.t. } \mathrm{E}_q \left[\boldsymbol{\beta} \right] \in \mathsf{C}$$

Assumptions:

- $\mathsf{C} \subset \mathcal{B}$ is closed.
- if q_* is a solution, then $\mathrm{E}_{q_*} \left[\boldsymbol{\beta} \right] = \mathbf{a}_*$ is bounded.

Representation result

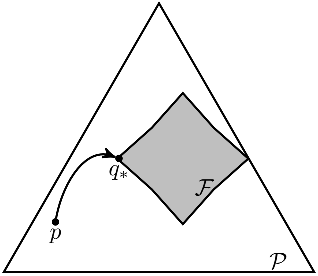
Every q_* can be represented as:

$$q_*(x) = p(x) e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x) \rangle - G(\boldsymbol{\lambda}_*)}$$

where $\boldsymbol{\lambda}_*$ is a function of $\mathbf{a}_*$.

Main Idea

Find solution using exponential family representation



$$\min_{\mathbf{c} \in \mathsf{A}} \left[\min_{q \in \mathcal{P}} \mathrm{KL}(q \| p) \text{ s.t. } \mathrm{E}_q \left[\boldsymbol{\beta} \right] = \mathbf{c} \right]$$

- Inner problem is "easy" to solve.
- Find compact $\mathsf{A} \subset \mathsf{C}$ so that $\mathbf{a}_* \in \mathsf{A}$.
- Define $\mathcal{F} = \{ q_{\mathbf{c}} \; \forall \; \mathbf{c} \in \mathsf{A} \} \subset \mathcal{P}$.
- Show that $q_* \in \mathcal{F}$.

Generalize to any $\mathcal{Q} \subset \mathcal{P}$ iff. $q_* \in \mathcal{Q}$.

Conjugate Priors for Relative Entropy

Motivation

- Exponential family representation requires normalization constant $G(\boldsymbol{\lambda})$.
- In many cases, computing $G(\boldsymbol{\lambda})$ is a challenging high dimensional problem.

Definition

The distribution $p \in \mathcal{Q}$ is a relative entropy conjugate prior distribution if any solution $q_* \in \mathcal{Q}$.

Conjugate Priors for Relative Entropy

Some properties

No need to compute $G(\boldsymbol{\lambda})$ explicitly if $p \in \mathcal{Q}$ is conjugate:

$$f_* = \operatorname*{arg\,min}_{f \in \mathcal{Q}} \mathrm{KL}(f \| p) \text{ s.t. } \mathrm{E}_f \left[\boldsymbol{\beta} \right] \in \mathsf{C}$$

and $f_* \equiv q_*$.

Proposition

Let

$$\mathcal{Q} = \left\{ p \;\middle|\; p_{\boldsymbol{\eta},\nu}(x) = h(x,\nu) e^{\langle \boldsymbol{\eta}, \boldsymbol{\beta}(x) \rangle - D(\boldsymbol{\eta},\nu)} \right\},$$

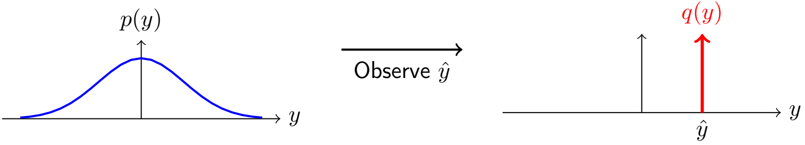
$\mathcal{Q}$ are a family of relative entropy conjugate priors i.e. any $q_* \in \mathcal{Q}$.

Constrained Bayesian Inference

Bayesian Inference $\subset$ Constrained Relative Entropy

Sample space: $\mathbb{X}$ latent variables and $\mathbb{Y}$ observations.

- Once you observe $y = \hat{y}$, there is no longer any uncertainty! Williams (1980)



- Hence $q(y) = \delta_{\hat{y}}$
- Equivalent constraint set $\mathrm{E}_q\left[\,\delta_{\hat{y}}\,\right] = 1$

Bayesian Inference ⊂ Constrained Relative Entropy

Solve:

$$\min_{q \in \mathcal{P}} \left[\mathrm{KL}(q(x,y) \| p(x,y)) \, \text{s.t.} \, \mathrm{E}_q \left[\, \delta_{\hat{y}} \, \right] = 1 \right]$$

Solution is the Bayesian posterior!

$$q_*(x,y) = q_*(x|y=\hat{y})\delta_{\hat{y}} = p(x|y=\hat{y})\delta_{\hat{y}}$$

In other words:

$$q(x|y=\hat{y}) \equiv p(x|y=\hat{y})$$

Bayesian Inference with Additional Constraints

Our approach - incorporate further constraints on $q(x)$.

Constrained Bayesian inference

Solve:

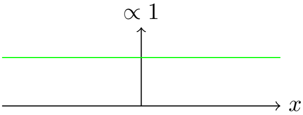
$$\min_{q(y),\ q(x|y)\in\mathcal{P}} \left[\mathrm{KL}(q(x|y)q(y)\|p(x,y)) \text{ s.t. } \substack{\mathrm{E}_q[\,\delta_{\hat{y}}\,]=1 \\ \mathrm{E}_q[\,\boldsymbol{\beta}(x)\,]\in\mathsf{C}} \right]$$

or the equivalent problem:

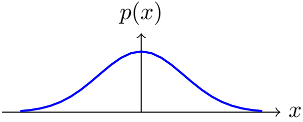
$$\min_{q\in\mathcal{P}} \left[\mathrm{KL}(q(x)\|p(x|y)) \text{ s.t. } \mathrm{E}_q\left[\,\boldsymbol{\beta}(x)\,\right] \in \mathsf{C} \right]$$

Table of Contents

Overview of the Approach

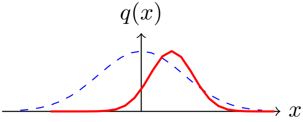


Domain $\mathbb{X}$
Complete ignorance



Maximize entropy given constraints

$$\min_p \left[\mathrm{H}(p) \ \textsf{s.t.}\ \mathrm{E}_q \left[\, \gamma(x) \,\right] \in \mathsf{D} \right]$$



Minimize relative entropy
given data and constraints

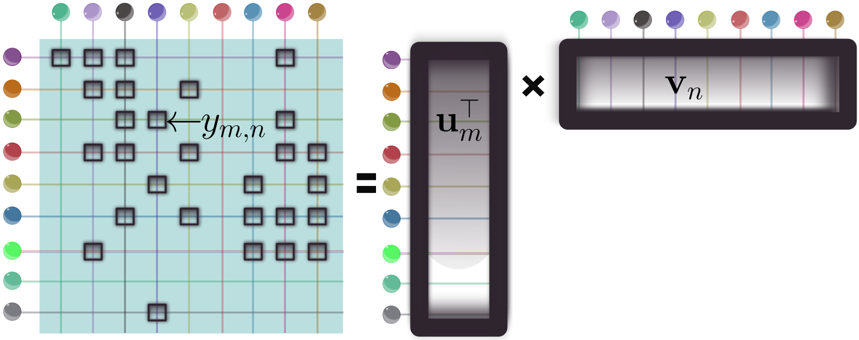
$$\min_q \left[\mathrm{KL}(q \| p) \ \textsf{s.t.}\ \mathrm{E}_q \left[\, \beta(x) \,\right] \in \mathsf{C} \right]$$

Illustrative Example

Low Rank Expectation Constraints

Matrix Factorization

- Estimate rank R factors $\mathbf{u}_m \in \mathbb{R}^R$ and $\mathbf{v}_n \in \mathbb{R}^R$ so that $y_{m,n} \approx \mathbf{u}_m^\top \mathbf{v}_n$.



Initial constraints and Prior Distribution

Domain:

- $\mathbf{y} \in \mathbb{R}^K$ with $(m, n) \in \mathsf{K}$
- $\mathbf{W} \in \mathbb{R}^{M \times N}$, $\mathbf{w} = \mathsf{vec}(\mathbf{W})$, let $\mathbf{P} : \mathsf{M} \times \mathsf{N} \mapsto \mathsf{K}$.

Initial information:

- $\mathrm{E}\left[\,\mathbf{y}|\mathbf{w}\,\right] = \mathbf{P}\mathbf{w}$, $\mathrm{E}\left[\,\mathbf{y}\mathbf{y}^\top\,\right] = \sigma^2\mathbf{I}$.
- $\mathrm{E}\left[\,\mathbf{w}\,\right] = \mathbf{0}$, $\mathrm{E}\left[\,\mathbf{w}\mathbf{w}^\top\,\right] = \mathbf{C} \otimes \mathbf{R}$.

Prior distribution

- Solve for maximum entropy distribution given constraints.
- Unique solution:

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}\left(\mathbf{P}\mathbf{w}, \sigma^2\right), \quad p(\mathbf{w}) = \mathcal{N}\left(\mathbf{0}, \mathbf{C} \otimes \mathbf{R}\right).$$

Constrained Relative Entropy Approach

- Observe $\mathbf{y} = \widehat{\mathbf{y}}$
- Additional constraints $\mathrm{rank}(\mathrm{E}\left[\,\mathbf{W}\,\right]) \leq R$.

Solution

- Let $\mathsf{C} = \{\mathbf{B} \,|\, \mathrm{rank}(\mathbf{B}) \leq R\}$
- $p(\mathbf{W}|\mathcal{D}) = \mathcal{N}\left(\mathbf{A}, \mathbf{\Sigma}\right)$ is the Bayesian posterior.
- p is a relative entropy conjugate prior for $\boldsymbol{\beta}(\mathbf{W}) = \mathbf{W}$.
- Hence q is Gaussian $q(\mathbf{W}) = \mathcal{N}\left(\mathbf{M}_*, \mathbf{S}_*\right)$ and:

$$\mathbf{M}_*, \mathbf{S}_* = \operatorname*{arg\,min}_{\mathbf{M},\mathbf{S}} \Big[\mathrm{KL}(\mathcal{N}\left(\mathbf{M}, \mathbf{S}\right) \| \mathcal{N}\left(\mathbf{A}, \mathbf{\Sigma}\right)) \text{ s.t. } \mathbf{M} \in \mathsf{C} \Big]$$

Bayesian Approach

Factor Model

Prior Distribution:

$$\mathbf{y}|\mathbf{w} \sim \mathcal{N}\left(\mathbf{P}\mathbf{w}, \sigma^2\right),$$
$$\mathbf{W} = \mathbf{U}\mathbf{V}^\top, \;\; \mathbf{u}_m = \mathcal{N}\left(\mathbf{0}, \mathbf{R}\right), \;\; \mathbf{v}_n = \mathcal{N}\left(\mathbf{0}, \mathbf{C}\right).$$

Inference:

- Easy to compute some marginals e.g. $p(\mathbf{y}|\mathbf{U})$ (Lawrence and Urtasun, 2009),
- but $p(\mathbf{U}, \mathbf{V}|\mathbf{y})$ is analytically intractable!

In practice:

- Compute MAP estimate (equivalent to matrix factorization)
- Variational approximation.
- Sampling: $\mathrm{E}_{p(\mathbf{U},\mathbf{V}|\mathbf{y})}\left[\,\mathbf{W}\,\right]$ is unlikely to be low rank.

Application to Multitask Learning

- Constrained multitask learning and parameter estimation:
  - Applied to recommender systems and fMRI.
  - Rank constraints on the postdata mean matrix.
  - Kronecker constraints on the postdata covariance matrix.
  - Sparsity constraints on prior precision matrix.
- Nonparametric multitask learning:
  - Applied to predictive modeling of transposable data.
  - Nuclear norm constraints on the mean function.
- Multitask bipartite ranking:
  - Applied to disease-gene prioritization.
  - Ordering constraints on ranking variables.
  - Nuclear norm constraints on mean function of regression variables.

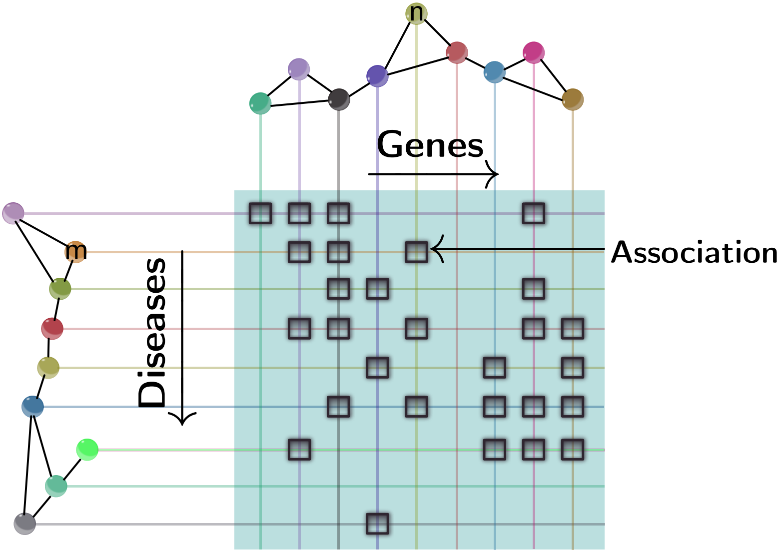
Multitask Bipartite Ranking

for Disease-Gene Prioritization

Genetic Diseases

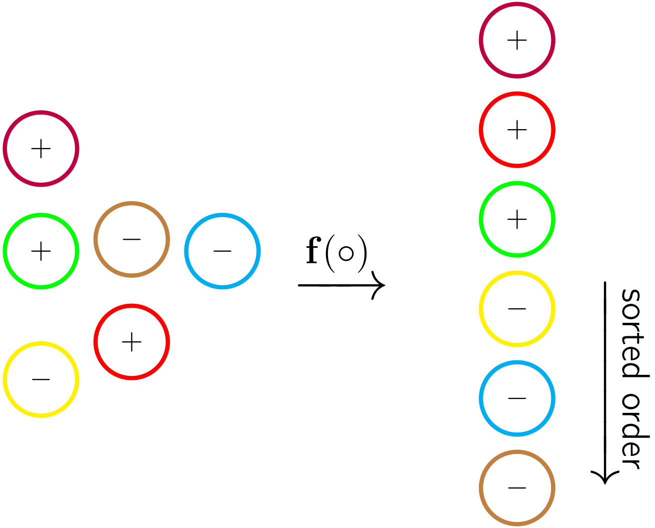
- Asthma, Cancer, Alzheimer's and many other diseases are associated with genetic mutations.
- Medical interest in unknown disease gene associations for targeted treatment, gene research, etc.
- $20,000 - 25,000$ human genes and thousands of diseases so enumerative search is computationally intractable.
- Computational models can be used to suggest most likely genetic associations.

Gene Disease Data



Bipartite Ranking

Inferring order from binary examples



Problem Setup

Sample Space:

- Associate $x_{m,n}$ to each observation $y_{m,n}$ (*ranking variable*).
- Associate $z_{m,n}$ to each $x_{m,n}$ (*regression variable*).

Prior Distribution

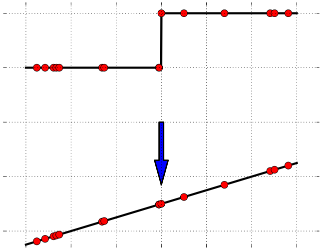
$$x_{m,n} \sim \mathcal{N}\left(z_{m,n}, \sigma^2\right), \quad Z \sim \mathcal{MGP}\left(0, \mathcal{C}_{\mathsf{M}}, \mathcal{C}_{\mathsf{N}}\right).$$

Regression variable constraint

Nuclear norm constraints for parsimony and performance:

$$\mathrm{E}\left[\,Z\,\right] \in \mathsf{T}_\epsilon = \{\Phi \mid \|\|\Phi\|\|_{1\text{-}\mathcal{H}_\mathcal{C}} \leq \epsilon\}$$

Ranking Variable Constraint



- Adapted from monotone retargeting (Acharyya et al., 2012)
- Key idea: Match ordering in $\mathbf{x}_m$ with ordering in $\mathbf{y}_m$
- Encoded as $\mathrm{E}\left[\,\mathbf{x}\,\right] \in \mathsf{R}_{\mathbf{y}}$

Inference

The resulting problem is given by:

$$q(\mathbf{x}, Z) = \arg\min_{q \in \mathcal{P}} \begin{bmatrix} & \mathrm{KL}(q(\mathbf{x}, Z) \| p(\mathbf{x}, Z)) \\ \text{s.t.} & \mathrm{E}_q\left[\,\mathbf{x}\,\right] \in \mathsf{R}_{\mathbf{y}},\ \mathrm{E}_q\left[\,Z\,\right] \in \mathsf{T}_\epsilon \end{bmatrix}$$

Solution

- $q(\mathbf{x}, Z)$ is a Gaussian process.
- Solution is unique!
- The marginal distribution $q(Z)$ is a matrix-varite GP with constrained (usually low rank) mean function.

Key Steps

- $q(\mathbf{x}, Z)$ is Gaussian for any finite index set via conjugacy.
- $\mathrm{KL}(q(\mathbf{x}, Z) \| p(\mathbf{x}, Z))$ reduces to $\mathrm{KL}(q(\mathbf{x}, \mathbf{Z}) \| p(\mathbf{x}, \mathbf{Z}))$
    - Proof uses representation theorem for spectral regularization (Abernethy et al., 2009).
    - $\mathbf{Z}$ is the finite "hidden" matrix corresponding to observations.
- Solve for $q(\mathbf{x}, \mathbf{Z})$.
- "Expand" finite $q(\mathbf{x}, \mathbf{Z})$ to random process $q(\mathbf{x}, Z)$
    - Using Kolmogorov's extension theorem (Kolmogorov, 1933).

Experiments

Datasets:

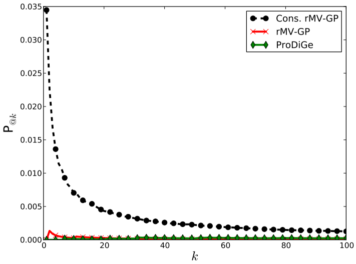
- **OMIM**: Set of manually-generated review articles linking inheritable diseases to one or a few genes ($M = 3,210,\ N = 13,614,\ J = 3,636,\ d = 0.0083\%$).
- **Medline**: Set of disease gene associations generated by processing annotations in the NIH PubMed/Medline literature ($M = 4,496,\ N = 21,243,\ J = 250,190,\ d = 0.36\%$).

Experimental Setup:

- "Known" diseases: randomly hidden associations.
- "New" diseases: randomly hidden diseases (*cold start*).

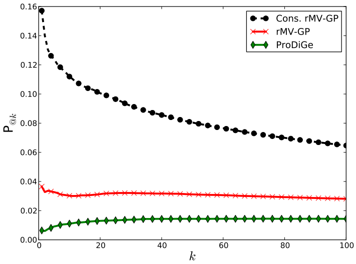
Results

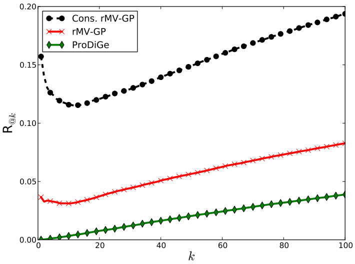
OMIM



Results

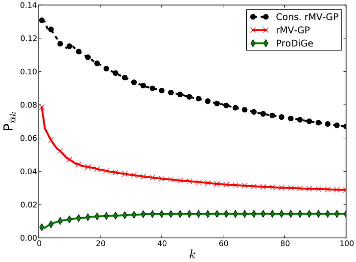
Medline "Known" Diseases

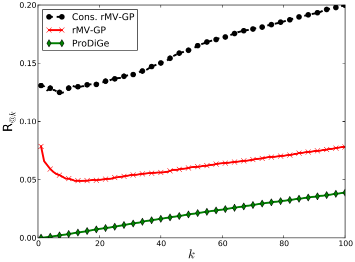




Results

Medline "New" Diseases





Predicted Candidate Genes

Alzheimer's Disease

Rank	Gene	Description
3	IGF1	Insulin-signaling protein linked to aging
9	MTHFR	Enzyme linked to cognitive impairments
15	IL1RN	Inflammatory marker linked to AD
16	APOE	Lipoprotein component linked to AD

Prostate Cancer

Rank	Gene	Description
2	CRP	Marker linked to onset age and survival
7	AGT	Downstream products linked to metastasis
9	VEGFA	Promotes blood vessel growth in tumors
13	MMP9	Protein linked to metastatic process

Table of Contents

Conclusion

Developed a framework for probabilistic inference subject to expectation constraints:

- constraints may be any combination of domain knowledge and data.
- constraint set may be nonconvex.

Applied the results to multitask learning:

- Constrained multitask learning and parameter estimation.
- Nonparametric multitask learning for transposable data.
- Multitask bipartite ranking for disease gene prioritization.

Future Work

- Constrained relative entropy inference is difficult to optimize using sampling techniques. These may be necessary for complicated problems.
- The presented inference approach can be extended to other divergence metrics such as Csiszár divergences and Bregman divergences.
- Hyperparameter optimization is the most challenging aspect of the current approach. More research is required to improve available methods for large scale problems.
- Further investigation is required to fully understand the biological implications of the multitask bipartite ranking model.

Thank You

Related Publications

- Oluwasanmi Koyejo and Joydeep Ghosh. Constrained Bayesian inference for low rank multitask learning. In *Proceedings of the 29th conference on Uncertainty in artificial intelligence (UAI)*, 2013b. (Oral)
- Oluwasanmi Koyejo and Joydeep Ghosh. A representation approach for relative entropy minimization with expectation constraints. In *Workshop on Divergences and Divergence Learning (WDDL)*, 2013a
- Cheng Lee, Oluwasanmi Koyejo, and Joydeep Ghosh. Identifying candidate disease genes using a trace norm constrained bipartite raking model. In *Engineering in Medicine and Biology Society (EMBC)*, 2013
- Sreangsu Acharyya*, Oluwasanmi Koyejo*, and Joydeep Ghosh. Learning to rank with Bregman divergences and monotone retargeting. In *Proceedings of the 28th conference on Uncertainty in artificial intelligence (UAI)*, 2012
- Oluwasanmi Koyejo and Joydeep Ghosh. A kernel-based approach to exploiting interaction-networks in heterogeneous information sources for improved recommender systems. In *HetRec '11*, pages 9–16, 2011

Related Publications

Submitted Publications

- Oluwasanmi Koyejo, Cheng Lee, and Joydeep Ghosh. Constrained relative entropy minimization with application to disease gene prioritization. 2013d. Under review
- Oluwasanmi Koyejo, Cheng Lee, and Joydeep Ghosh. A constrained matrix-variate Gaussian process for transposable data. 2013c. Under review
- Sreangsu Acharyya, Oluwasanmi Koyejo, and Joydeep Ghosh. Learning to rank with Bregman divergences and monotone retargeting. 2013. Under review

Other Publications

- Mijung Park*, Oluwasanmi Koyejo*, Joydeep Ghosh, Russell R. Poldrack, and Jonathan W. Pillow. Bayesian structure learning for functional neuroimaging. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013
- O. Koyejo, P. Patel, J. Ghosh, and R. A. Poldrack. Learning predictive cognitive structure from fMRI using supervised topic models. In *International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, 2013a
- Oluwasanmi Koyejo, Cheng Lee, and Joydeep Ghosh. The trace norm constrained matrix-variate Gaussian process for multitask bipartite ranking. *CoRR*, abs/1302.2576, 2013e
- Oluwasanmi Koyejo and Joydeep Ghosh. MiPPS; a generative model for multi-manifold clustering. In *AAAI Fall Symposium on Manifold Learning and Its Applications*. AAAI Press, 2009
- Oluwasanmi Koyejo. Manifold learning and its applications: Reports of the AAAI 2010 fall symposia. *AI Magazine*, 32(1):93–100, 2011
- Richard Souvenir and Oluwasanmi Koyejo. Manifold learning and its applications: Reports of the AAAI 2009 fall symposia. *AI Magazine*, 31 (1):88–94, 2010

Other Publications

Submitted Publications

- R. A. Poldrack, D. M. Barch, J. P. Mitchell, T. D. Wager, A. D. Wagner, J. T. Devlin, C. Cumba, O. Koyejo, and M. P. Milham. Towards open sharing of task-based fMRI data: The OpenfMRI project. 2013. Under review
- Oluwasanmi Koyejo, Sreangsu Acharyya, and Joydeep Ghosh. Retargeted matrix factorization. 2013b. Under review

References I

Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *JMLR*, 10:803–826, 2009. ISSN 1532-4435.

Sreangsu Acharyya, Oluwasanmi Koyejo, and Joydeep Ghosh. Learning to rank with bregman divergences and monotone retargeting. In *Proceedings of the 28th conference on Uncertainty in artificial intelligence*, UAI '12, 2012. ISBN 0-9749039-0-6.

Sreangsu Acharyya*, Oluwasanmi Koyejo*, and Joydeep Ghosh. Learning to rank with Bregman divergences and monotone retargeting. In *Proceedings of the 28th conference on Uncertainty in artificial intelligence (UAI)*, 2012.

Sreangsu Acharyya, Oluwasanmi Koyejo, and Joydeep Ghosh. Learning to rank with Bregman divergences and monotone retargeting. 2013. Under review.

Yasemin Altun and Alexander J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March 1996. ISSN 0891-2017.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *J. Mach. Learn. Res.*, 8:1217–1260, December 2007. ISSN 1532-4435.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *NIPS*. MIT Press, 1999.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review Online Archive (Prola)*, 106(4):620–630, 1957.

A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea, New York, 1933. Trans. N. Morrison (1956).

O. Koyejo, P. Patel, J. Ghosh, and R. A. Poldrack. Learning predictive cognitive structure from fMRI using supervised topic models. In *International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, 2013a.

References II

Oluwasanmi Koyejo. Manifold learning and its applications: Reports of the AAAI 2010 fall symposia. *AI Magazine*, 32(1):93–100, 2011.

Oluwasanmi Koyejo and Joydeep Ghosh. MiPPS; a generative model for multi-manifold clustering. In *AAAI Fall Symposium on Manifold Learning and Its Applications*. AAAI Press, 2009.

Oluwasanmi Koyejo and Joydeep Ghosh. A kernel-based approach to exploiting interaction-networks in heterogeneous information sources for improved recommender systems. In *HetRec '11*, pages 9–16, 2011.

Oluwasanmi Koyejo and Joydeep Ghosh. A representation approach for relative entropy minimization with expectation constraints. In *Workshop on Divergences and Divergence Learning (WDDL)*, 2013a.

Oluwasanmi Koyejo and Joydeep Ghosh. Constrained Bayesian inference for low rank multitask learning. In *Proceedings of the 29th conference on Uncertainty in artificial intelligence (UAI)*, 2013b. (Oral).

Oluwasanmi Koyejo, Sreangsu Acharyya, and Joydeep Ghosh. Retargeted matrix factorization. 2013b. Under review.

Oluwasanmi Koyejo, Cheng Lee, and Joydeep Ghosh. A constrained matrix-variate Gaussian process for transposable data. 2013c. Under review.

Oluwasanmi Koyejo, Cheng Lee, and Joydeep Ghosh. Constrained relative entropy minimization with application to disease gene prioritization. 2013d. Under review.

Oluwasanmi Koyejo, Cheng Lee, and Joydeep Ghosh. The trace norm constrained matrix-variate Gaussian process for multitask bipartite ranking. *CoRR*, abs/1302.2576, 2013e.

Solomon Kullback. *Information Theory and Statistics*. Dover, 1959.

Neil D. Lawrence and Raquel Urtasun. Non-linear matrix factorization with gaussian processes. In *ICML*, pages 601–608. ACM, 2009.

Cheng Lee, Oluwasanmi Koyejo, and Joydeep Ghosh. Identifying candidate disease genes using a trace norm constrained bipartite raking model. In *Engineering in Medicine and Biology Society (EMBC)*, 2013.

References III

Mijung Park*, Oluwasanmi Koyejo*, Joydeep Ghosh, Russell R. Poldrack, and Jonathan W. Pillow. Bayesian structure learning for functional neuroimaging. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

R. A. Poldrack, D. M. Barch, J. P. Mitchell, T. D. Wager, A. D. Wagner, J. T. Devlin, C. Cumba, O. Koyejo, and M. P. Milham. Towards open sharing of task-based fMRI data: The OpenfMRI project. 2013. Under review.

Richard Souvenir and Oluwasanmi Koyejo. Manifold learning and its applications: Reports of the AAAI 2009 fall symposia. *AI Magazine*, 31(1):88–94, 2010.

Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

Jun Zhu. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.

Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *ICML*, 2009.