Quasi-Bayesian Learning

An application to NMF

Benjamin Guedj

https://bguedj.github.io
Inria Lille - Nord Europe

TAU

Batch Learning in a Nutshell

Collect a sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d replications of some random variable $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$.

Goal: use $\mathcal{D}_n$ to build up $\widehat{\phi}$ such that $\widehat{\phi}(\mathbf{X})$ is an "acceptable" prediction of $\mathbf{Y}$.

For some loss function ℓ, let

$$R \colon \widehat{\phi} \mapsto \mathbb{E}\ell\left(\widehat{\phi}(\mathbf{X}), \mathbf{Y}\right) \quad \text{and} \quad R_n \colon \widehat{\phi} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{\phi}(\mathbf{X}_i), \mathbf{Y}_i\right)$$

denote the risk and empirical risk, respectively.

The quasi-Bayesian approach

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

For some (inverse temperature) parameter $\lambda > 0$, quasi-posterior

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right)\pi(\cdot).$$

The quasi-Bayesian approach

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

For some (inverse temperature) parameter $\lambda > 0$, quasi-posterior

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right) \pi(\cdot).$$

In general, $\exp\left(-\lambda R_n(\cdot)\right)$ is not a likelihood (hence the term quasi-Bayesian).

A variational perspective

With the classical quadratic loss $\ell\colon (a, b) \mapsto (a - b)^2$,

$$\widehat{\rho}_\lambda \in \operatorname*{arg\,inf}_{\rho \ll \pi} \left\{ \int_{\mathcal{F}} R_n(\phi)\rho(\mathrm{d}\phi) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\},$$

where $\mathcal{K}$ is the Kullback-Leibler divergence

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int_{\mathcal{F}} \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho & \text{when } \rho \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Typical quasi-Bayesian estimators

MAQP

$$\widehat{\phi}_\lambda \in \mathop{\mathsf{arg\,max}}_{\phi \in \mathcal{F}} \widehat{\rho}_\lambda(\phi).$$

Mean

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \widehat{\rho}_\lambda(\mathrm{d}\phi).$$

Realization

$$\widehat{\phi}_\lambda \sim \widehat{\rho}_\lambda.$$

And so on.

Statistical aggregation revisited

Assume that $\mathcal{F}$ is finite.

Statistical aggregation revisited

Assume that $\mathcal{F}$ is finite.

The mean of the quasi-posterior $\widehat{\rho}_\lambda$ amounts to the celebrated exponentially weighted aggregate (EWA)

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} \phi = \sum_{i=1}^{\#\mathcal{F}} \omega_{\lambda,i} \phi_i$$

where

$$\omega_{\lambda,i} = \frac{\exp(-\lambda R_n(\phi_i))\pi(\phi_i)}{\sum_{j=1}^{\#\mathcal{F}} \exp(-\lambda R_n(\phi_j))\pi(\phi_j)}.$$

G. (2013). Agrégation d'estimateurs et de classificateurs : théorie et méthodes, *Ph.D. thesis, Université Pierre & Marie Curie*

Probably Approximately Correct (PAC) oracle inequalities

Let $R^\star$ denote the Bayes risk and set $\lambda \propto n$. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$.

Key argument: concentration inequalities (*e.g.*, Bernstein) + duality formula (Csiszár, Catoni).

Probably Approximately Correct (PAC) oracle inequalities

Let $R^\star$ denote the Bayes risk and set $\lambda \propto n$. For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha}\right\}\right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$.

Key argument: concentration inequalities (*e.g.*, Bernstein) + duality formula (Csiszár, Catoni).

Typical regimes in the literature

- $\alpha = \frac{1}{2}$ (slow rate)
- $\alpha = 1$ (fast rate)

$d := \dim(\mathcal{X})$

- $\Delta(\phi, \epsilon) \propto d + \log \frac{1}{\epsilon}$
- $\Delta(\phi, \epsilon) \propto \log d + \log \frac{1}{\epsilon}$

Lemma (Catoni, 2004)

Let $(A, \mathcal{A})$ be a measurable space. For any probability μ on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int (\exp \circ h) \mathrm{d}\mu < \infty$,

$$\log \int (\exp \circ h) \mathrm{d}\mu = \sup_{m \in \mathcal{M}_\pi(A, \mathcal{A})} \left\{ \int h \mathrm{d}m - \mathcal{K}(m, \mu) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to m on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp \circ h(a)}{\int (\exp \circ h) \mathrm{d}\mu}, \quad a \in A.$$

The PAC-Bayesian theory

The PAC-Bayesian theory

...consists in producing PAC inequalities of Bayesian-flavored (such as quasi-Bayesian) estimators.

The PAC-Bayesian theory

...consists in producing PAC inequalities of Bayesian-flavored (such as quasi-Bayesian) estimators.

Shawe-Taylor and Williamson (1997). A PAC analysis of a Bayes estimator, *COLT*

McAllester (1998). Some PAC-Bayesian theorems, *COLT*

McAllester (1999). PAC-Bayesian model averaging, *COLT*

Catoni (2004). Statistical Learning Theory and Stochastic Optimization, Springer

Audibert (2004). Une approche PAC-bayésienne de la théorie statistique de l'apprentissage, *Ph.D. thesis, Université Pierre & Marie Curie*

Catoni (2007). PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, IMS

Dalalyan and Tsybakov (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity, *Machine Learning*

A flexible and powerful framework

A flexible and powerful framework

Numerous models addressed by the PAC-Bayes literature

📘 Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*

📘 Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*

📘 Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

📘 G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

📘 G. and Robbiano (2015). PAC-Bayesian High Dimensional Bipartite Ranking, *arXiv preprint*

📘 Li, G. and Loustau (2016). A Quasi-Bayesian perspective to Online Clustering, *arXiv preprint*

📘 Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

A flexible and powerful framework

Numerous models addressed by the PAC-Bayes literature

Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*

Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*

Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

G. and Robbiano (2015). PAC-Bayesian High Dimensional Bipartite Ranking, *arXiv preprint*

Li, G. and Loustau (2016). A Quasi-Bayesian perspective to Online Clustering, *arXiv preprint*

Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

Towards (almost) no assumptions to derive powerful results

Bégin, Germain, Laviolette and Roy (2016). PAC-Bayesian bounds based on the Rényi divergence, *AISTATS*

Alquier and G. (2016). Simpler PAC-Bayesian bounds for hostile data, *arXiv preprint*

(PAC inequalities for heavy-tailed time series)

In practice...

Previous instantiations of $\widehat{\phi}_\lambda$ are not tractable.

Instead of an infinite-dimensional functional space $\mathcal{F}$, we often resort to some projection onto $\mathbb{R}^d$.

Sampling from a d-dimensional non-standard distribution is still an algorithmic challenge.

Existing implementation

- (Transdimensional) MCMC

  G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

  Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

  G. and Robbiano (2015). PAC-Bayesian High Dimensional Bipartite Ranking, *arXiv preprint*

  Li, G. and Loustau (2016). A Quasi-Bayesian perspective to Online Clustering, *arXiv preprint*

Existing implementation

- (Transdimensional) MCMC

  G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

  Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

  G. and Robbiano (2015). PAC-Bayesian High Dimensional Bipartite Ranking, *arXiv preprint*

  Li, G. and Loustau (2016). A Quasi-Bayesian perspective to Online Clustering, *arXiv preprint*

- Stochastic optimization

  Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

Existing implementation

- (Transdimensional) MCMC

  📘 G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

  📘 Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

  📘 G. and Robbiano (2015). PAC-Bayesian High Dimensional Bipartite Ranking, *arXiv preprint*

  📘 Li, G. and Loustau (2016). A Quasi-Bayesian perspective to Online Clustering, *arXiv preprint*

- Stochastic optimization

  📘 Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

- Variational Bayes

  📘 Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors, *Journal of Machine Learning Research*

Bridging the gap between theory and implementation

Goal: PAC oracle inequalities for approximations of $\widehat{\rho}_\lambda$ (echoes the celebrated statistical / computational tradeoff).

Bridging the gap between theory and implementation

Goal: PAC oracle inequalities for approximations of $\widehat{\rho}_\lambda$ (echoes the celebrated statistical / computational tradeoff).

Let $\widetilde{\rho}_\lambda$ denote a VB approximation of $\widehat{\rho}_\lambda$. The rate of convergence in PAC inequalities is of analogous magnitude for $\widetilde{\rho}_\lambda$ and $\widehat{\rho}_\lambda$.

Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors, *Journal of Machine Learning Research*

Bridging the gap between theory and implementation

Goal: PAC oracle inequalities for approximations of $\widehat{\rho}_\lambda$ (echoes the celebrated statistical / computational tradeoff).

Let $\widetilde{\rho}_\lambda$ denote a VB approximation of $\widehat{\rho}_\lambda$. The rate of convergence in PAC inequalities is of analogous magnitude for $\widetilde{\rho}_\lambda$ and $\widehat{\rho}_\lambda$.

Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors, *Journal of Machine Learning Research*

MCMC for online (sequential) quasi-Bayesian learning: the stationary distribution of the Markov Chain is indeed $\widehat{\rho}_\lambda$.

Li, G. and Loustau (2016). A Quasi-Bayesian perspective to Online Clustering, *arXiv preprint*

Quasi-Bayesian Non-Negative Matrix Factorization

Alquier and G. (2017)

An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization

Mathematical Methods of Statistics

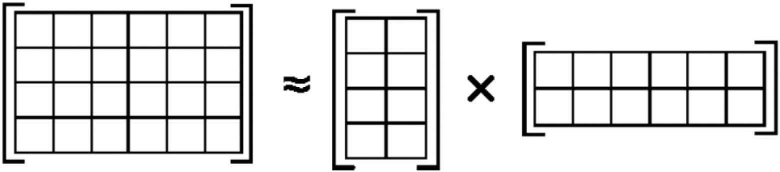
NMF

NMF amounts to decompose an $m_1 \times m_2$ matrix M as a product of two low rank matrices with non-negative entries.

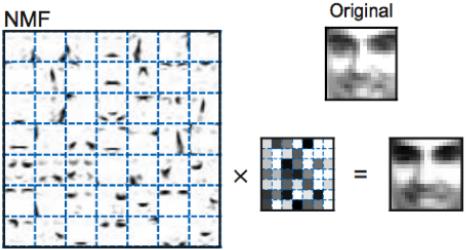
$$M \simeq UV^\top,$$

where U is $m_1 \times K$ and V is $m_2 \times K$, and $K \ll m_1 \wedge m_2$.

$M_{\cdot,j} \simeq \sum_{\ell=1}^{K} V_{j,\ell} U_{\cdot,\ell}$.



Wide range of applications (image processing, separation of sources in audio and video files, topics extraction in text, recommender systems...)



Separation of audio sources [Demo, courtesy of C. Févotte]

Setting

We observe an $m_1 \times m_2$ matrix Y and we assume

$$Y = M + \mathcal{E}$$

with $\mathbb{E}(\mathcal{E}) = 0$ and $\mathbb{V}(\mathcal{E}) = \sigma^2 \mathrm{Id}$.

Our goal is to find a "good" factorization of M.

Notation

Frobenius norm

$$\|A\|_F = \sqrt{\langle A, A \rangle_F},$$

$$\langle A, B \rangle_F = \mathrm{Tr}(AB^\top) = \sum_{i=1}^{p} \sum_{j=1}^{q} A_{i,j} B_{i,j}.$$

For any $r \in \{1, \ldots, K\}$, $\mathcal{M}_r(L)$ is the set of matrices U^0 with non-negative entries bounded by L such that

$$U^0 = \begin{pmatrix} U^0_{11} & \ldots & U^0_{1r} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ U^0_{m_1 1} & \ldots & U^0_{m_1 r} & 0 & \ldots & 0 \end{pmatrix}$$

Assumption

The entries of $\mathcal{E}$ are i.i.d and $\mathbb{E}\mathcal{E}_{i,j} = 0$. Let $m(x) = \mathbb{E}[\mathcal{E}_{i,j}\mathbb{1}_{\mathcal{E}_{i,j}\leq x}]$ and $F(x) = \mathbb{P}(\mathcal{E}_{i,j} \leq x)$.

There exists a nonnegative and bounded function g such that $\|g\|_\infty \leq 1$ and

$$\int_u^v m(x)\mathrm{d}x = \int_u^v g(x)\mathrm{d}F(x).$$

Assumption

The entries of $\mathcal{E}$ are i.i.d and $\mathbb{E}\mathcal{E}_{i,j} = 0$. Let $m(x) = \mathbb{E}[\mathcal{E}_{i,j} \mathbb{1}_{\mathcal{E}_{i,j} \leq x}]$ and $F(x) = \mathbb{P}(\mathcal{E}_{i,j} \leq x)$.

There exists a nonnegative and bounded function g such that $\|g\|_\infty \leq 1$ and

$$\int_u^v m(x)\mathrm{d}x = \int_u^v g(x)\mathrm{d}F(x).$$

This assumption is met whenever $\mathcal{E}_{i,j} \sim \mathcal{N}(0, \sigma^2)$ ($\|g\|_\infty = \sigma^2$) or $\mathcal{E}_{i,j} \sim \mathcal{U}(-b, b)$ ($\|g\|_\infty = b^2/2$).

Prior

For any a, $x > 0$, $g_a(x) = \frac{1}{a} f\left(\frac{x}{a}\right)$.

$$\forall \ell = 1, \dots, K, \quad \gamma_\ell \overset{\text{ind.}}{\sim} h,$$

$$\forall i = 1, \dots, m_1, j = 1, \dots, m_2, \quad U_{i,\ell}, V_{j,\ell} \overset{\text{ind.}}{\sim} g_{\gamma_\ell},$$

$$\pi(U, V, \gamma) = \prod_{\ell=1}^{K} \left(\prod_{i=1}^{m_1} g_{\gamma_\ell}(U_{i,\ell}) \right) \left(\prod_{j=1}^{m_2} g_{\gamma_\ell}(V_{j,\ell}) \right) h(\gamma_\ell),$$

and

$$\pi(U, V) = \int_{\mathbb{R}_+^K} \pi(U, V, \gamma) \mathrm{d}\gamma.$$

Prior (continued)

The idea behind this prior is that under h, many γ_ℓ should be small and lead to non-significant columns $U_{\cdot,\ell}$ and $V_{\cdot,\ell}$ (sufficient probability mass for h, around zero and elsewhere).

This is achieved by assuming[1]

1. $\exists\ 0 < \alpha < 1$, $\beta \geq 0$ and $\delta > 0$ such that for any $0 < \epsilon \leq \frac{1}{2\sqrt{2}S_f}$,

$$\int_0^\epsilon h(x)\mathrm{d}x \geq \alpha\epsilon^\beta \qquad \text{and} \qquad \int_1^2 h(x)\mathrm{d}x \geq \delta.$$

2. $\exists$ a non-increasing density $\widetilde{f}$ and $C > 0$ such that for any $x > 0$, $f(x) \geq C\widetilde{f}(x)$.

[1] $S_f := \max\left(1, \int_0^\infty x^2 f(x)\mathrm{d}x\right)$

Popular choices for f:

1. Exponential prior $f(x) = \exp(-x)$.
2. Truncated Gaussian prior $f(x) \propto \exp(2ax - x^2)$ with $a \in \mathbb{R}$.
3. Heavy-tailed prior $f(x) \propto \frac{1}{(1+x)^\zeta}$ with $\zeta > 1$.

The heavier the tails, the better the performance of QBNMF. But computational cost arises!

Popular choices for h:

1. Uniform distribution on $[0, c]$.
2. Inverse gamma prior $h(x) = \frac{b^a}{\Gamma(a)} \frac{1}{x^{a+1}} \exp\left(-\frac{b}{x}\right)$.
3. Gamme $\Gamma(a, b)$ prior for $a, b > 0$.

Quasi-Bayesian estimator

$$\widehat{\rho}_\lambda(U, V, \gamma) = \frac{1}{Z} \exp\left[-\lambda \|Y - UV^\top\|_F^2\right] \pi(U, V, \gamma),$$

where

$$Z := \int \exp\left[-\lambda \|Y - UV^\top\|_F^2\right] \pi(U, V, \gamma) \mathrm{d}(U, V, \gamma).$$

Quasi-Bayesian estimator

$$\widehat{\rho}_\lambda(U, V, \gamma) = \frac{1}{Z} \exp\left[-\lambda \|Y - UV^\top\|_F^2\right] \pi(U, V, \gamma),$$

where

$$Z := \int \exp\left[-\lambda \|Y - UV^\top\|_F^2\right] \pi(U, V, \gamma) \mathrm{d}(U, V, \gamma).$$

$$\widehat{M}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} UV^T = \int UV^T \widehat{\rho}_\lambda(U, V, \gamma) \mathrm{d}(U, V, \gamma).$$

Bayesian $\subset$ Quasi-Bayesian ($\subset$ PAC-Bayesian)

Bayesian $\subset$ Quasi-Bayesian ($\subset$ PAC-Bayesian)

The specific choice $\mathcal{E}_{i,j} \sim \mathcal{N}(0, 1/(2\lambda))$ (or rather, $\mathcal{E}_{i,j} \sim \mathcal{N}(0, \sigma^2)$ and $\lambda = 1/(2\sigma^2)$) turns our procedure fully Bayesian!

In this case the likelihood is written with the Frobenius norm, acting as a fitting criterion (other choices in the literature: Poisson likelihood, Itakura-Saito divergence).

Main result: sharp oracle inequality (simplified)

Fix $\lambda = 1/4$.

$$\mathbb{E}\left(\|\widehat{M}_\lambda - M\|_F^2\right) \leq \inf_{1\leq r\leq K} \inf_{(U^0,V^0)\in\mathcal{M}_r(L)} \Bigg\{ \|U^0 V^{0\top} - M\|_F^2 + r\left[8(m_1 \vee m_2)\log\left(\frac{2(L+1)^2 m_1 m_2}{C\widetilde{f}(L+1)}\right) + 8 + \log\frac{1}{\delta}\right] + K\left[4\beta\log\left(2S_f(L+1)^2 m_1 m_2\right) + 4\log\frac{1}{\alpha}\right]\Bigg\} + 4\log 4.$$

Main result: sharp oracle inequality (simplified)

Fix $\lambda = 1/4$.

$$\mathbb{E}\left(\|\widehat{M}_\lambda - M\|_F^2\right) \leq \inf_{1\leq r\leq K} \inf_{(U^0,V^0)\in\mathcal{M}_r(L)} \left\{ \|U^0 V^{0\top} - M\|_F^2 \right.$$

$$+ r\left[8(m_1 \vee m_2)\log\left(\frac{2(L+1)^2 m_1 m_2}{C\widetilde{f}(L+1)}\right) + 8 + \log\frac{1}{\delta}\right]$$

$$\left. + K\left[4\beta\log\left(2S_f(L+1)^2 m_1 m_2\right) + 4\log\frac{1}{\alpha}\right]\right\} + 4\log 4.$$

$$r(m_1 \vee m_2)\log\left(\frac{L^2 m_1 m_2}{C\widetilde{f}(L+1)}\right) = \begin{cases} r(m_1 \vee m_2)\log(m_1 m_2) & \text{if } L^2 = \mathcal{O}(1), \\ r(m_1 \vee m_2)L^2\log(Lm_1 m_2) & \text{if } f(x) \propto \exp(2ax - x^2) \\ r(m_1 \vee m_2)(\zeta+2)\log(Lm_1 m_2) & \text{if } f(x) \propto (1+x)^{-\zeta} \end{cases}$$

Gibbs sampler

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$.

For $k = 1, \ldots, N$:

Gibbs sampler

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$.

For $k = 1, \ldots, N$:

- **For** $i = 1, \ldots, m_1$: draw
  $U_{i,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(U_{i,\cdot} | V^{(k-1)}, \gamma^{(k-1)}, Y)$.
- **For** $j = 1, \ldots, m_2$: draw
  $V_{j,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(V_{j,\cdot} | U^{(k)}, \gamma^{(k-1)}, Y)$.
- **For** $\ell = 1, \ldots, K$: draw
  $\gamma_\ell^{(k)} \sim \widehat{\rho}_\lambda(\gamma_\ell | U^{(k)}, V^{(k)}, Y)$.

Gibbs sampler

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$.

For $k = 1, \ldots, N$:

- **For** $i = 1, \ldots, m_1$: draw
  $U_{i,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(U_{i,\cdot} | V^{(k-1)}, \gamma^{(k-1)}, Y)$.
- **For** $j = 1, \ldots, m_2$: draw
  $V_{j,\cdot}^{(k)} \sim \widehat{\rho}_\lambda(V_{j,\cdot} | U^{(k)}, \gamma^{(k-1)}, Y)$.
- **For** $\ell = 1, \ldots, K$: draw
  $\gamma_\ell^{(k)} \sim \widehat{\rho}_\lambda(\gamma_\ell | U^{(k)}, V^{(k)}, Y)$.

For the exponential prior, $\widehat{\rho}_\lambda(U_{i,\cdot} | V, \gamma, Y)$ amounts to a truncated Gaussian distribution.

Block coordinate descent

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$.

While not converged, $k := k + 1$:

Block coordinate descent

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$.

While not converged, $k := k + 1$:

$$U^{(k)} := \arg\min_{U} \left\{ \lambda \| Y - U(V^{(k-1)})^\top \|_F^2 - \sum_{i=1}^{m_1} \sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(U_{i,\ell})] \right\}$$

$$V^{(k)} := \arg\min_{V} \left\{ \lambda \| Y - U^{(k)} V^\top \|_F^2 - \sum_{j=1}^{m_2} \sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(V_{j,\ell})] \right\}$$

$$\gamma^{(k)} := \arg\min_{\gamma} \sum_{\ell=1}^{K} \left\{ - \sum_{i=1}^{m_1} \log[g_{\gamma_\ell}(U_{i,\ell}^{(k)})] - \sum_{j=1}^{m_2} \log[g_{\gamma_\ell}(V_{j,\ell}^{(k)})] - \log[h(\gamma_\ell)] \right\}$$

Block coordinate descent

Input Y, λ.

Initialization $U^{(0)}$, $V^{(0)}$, $\gamma^{(0)}$.

While not converged, $k := k+1$:

$$U^{(k)} := \arg\min_{U} \left\{ \lambda \| Y - U (V^{(k-1)})^\top \|_F^2 - \sum_{i=1}^{m_1} \sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(U_{i,\ell})] \right\}$$

$$V^{(k)} := \arg\min_{V} \left\{ \lambda \| Y - U^{(k)} V^\top \|_F^2 - \sum_{j=1}^{m_2} \sum_{\ell=1}^{K} \log[g_{\gamma_\ell^{(k-1)}}(V_{j,\ell})] \right\}$$

$$\gamma^{(k)} := \arg\min_{\gamma} \sum_{\ell=1}^{K} \left\{ - \sum_{i=1}^{m_1} \log[g_{\gamma_\ell}(U^{(k)}_{i,\ell})] - \sum_{j=1}^{m_2} \log[g_{\gamma_\ell}(V^{(k)}_{j,\ell})] - \log[h(\gamma_\ell)] \right\}$$

Public python library + demo USPS data (LeCun et al., 1990)

Take-home messages

- {Quasi,PAC}-Bayesian learning is a flexible and powerful machinery.

Take-home messages

- {Quasi,PAC}-Bayesian learning is a flexible and powerful machinery.

- First sharp oracle inequality in the literature for (QB-)NMF, showing adaptation to the rank.

Shameless self-promotion

Shameless self-promotion

NIPS 2017 Workshop

(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights

Long Beach Convention Center, California

December 9, 2017

Shameless self-promotion

What this talk could have been about: online clustering, high-dimensional ranking, PAC-Bayesian bounds for hostile data, stability, sequential principal curves...

Shameless self-promotion

NIPS 2017 Workshop

(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights

Long Beach Convention Center, California
December 9, 2017

What this talk could have been about: online clustering, high-dimensional ranking, PAC-Bayesian bounds for hostile data, stability, sequential principal curves...

Ongoing projects:
{active, agnostic/objective, deep, representation} learning (mostly with some PAC-Bayes)

https://bguedj.github.io

https://bguedj.github.io/nips2017/50shadesbayesian.html