Exponential Family Matrix Completion under Structural Constraints

Suriya Gunasekar SURIYA@UTEXAS.EDU
Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
The University of Texas at Austin, Texas, USA

Abstract

We consider the matrix completion problem of recovering a structured matrix from noisy and partial measurements. Recent works have proposed tractable estimators with strong statistical guarantees for the case where the underlying matrix is low–rank, and the measurements consist of a subset, either of the exact individual entries, or of the entries perturbed by additive Gaussian noise, which is thus implicitly suited for thin–tailed continuous data. Arguably, common applications of matrix completion require estimators for (a) heterogeneous data–types, such as skewed–continuous, count, binary, etc., (b) for heterogeneous noise models (beyond Gaussian), which capture varied uncertainty in the measurements, and (c) heterogeneous structural constraints beyond low–rank, such as block–sparsity, or a superposition structure of low–rank plus elementwise sparseness, among others. In this paper, we provide a vastly unified framework for generalized matrix completion by considering a matrix completion setting wherein the matrix entries are sampled from any member of the rich family of *exponential family distributions*; and impose general structural constraints on the underlying matrix, as captured by a general regularizer $\mathcal{R}(.)$. We propose a simple convex regularized M–estimator for the generalized framework, and provide a unified and novel statistical analysis for this general class of estimators. We finally corroborate our theoretical results on simulated datasets.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

1. Introduction

In the general problem of matrix completion, we seek to recover a structured matrix from noisy and partial measurements. This problem class encompasses a wide range of practically important applications such as recommendation systems, recovering gene–protein interactions, and analyzing document collections in language processing, among others. In recent years, leveraging developments in sparse estimation and compressed sensing, there has been a surge of work on computationally tractable estimators with strong statistical guarantees, specifically for the setting where a subset of entries of a low–rank matrix are observed either deterministically, or perturbed by additive noise that is Gaussian (Candes & Plan, 2010), or more generally sub–Gaussian (Keshavan et al., 2010b; Negahban & Wainwright, 2012). While such a Gaussian noise model is amenable to the subtle statistical analyses required for the ill–posed problem of matrix completion, it is not always practically suitable for all data settings encountered in matrix completion problems. For instance, such a Gaussian error model might not be appropriate in recommender systems based on movie ratings that are either binary (likes or dislikes), or range over the integers one through five. The *first question* we ask in this paper is whether we can generalize the statistical estimators for matrix completion as well as their analyses to general noise models? Note that a noise model captures the uncertainty underlying the matrix measurements, and is thus an important component of the problem specification given any application; and it is thus vital for broad applicability of the class of matrix completion estimators to extend to general noise models.

Though this might seem like a narrow technical, although important question, it is related to a broader issue. A Gaussian observation model implicitly assumes the matrix values are continuous–valued (and that they are thin–tail–distributed). But in modern applications, matrix data span the gamut of heterogeneous data–types, for instance, skewed–continuous, categorical–discrete including binary, count–valued, among others. This, thus gives rise to the *second question* of whether we can generalize the stan-

dard matrix completion estimators and statistical analyses, suited for thin–tailed continuous data, to more heterogeneous data–types? Note that there has been some recent work for the specific case of binary data by Davenport et al. (2012), but generalizations to other data–types and distributions is largely unexplored.

The recent line of work on matrix completion, moreover, enforces the constraint that the underlying matrix be either exactly or approximately low–rank. Aside from the low–rank constraints, further assumptions to eliminate overly "spiky" matrices are required for well–posed recovery under partial measurements (Candes & Recht, 2009). Early work provided generalization error bounds for various low–rank matrix completion algorithms, including algorithms based on nuclear norm minimization (Candes & Recht, 2009; Candes & Tao, 2010; Candes & Plan, 2010; Recht, 2011), max–margin matrix factorization (Srebro et al., 2004), spectral algorithms (Keshavan et al., 2010a;b), and alternating minimization (Jain et al., 2013). These work made stringent matrix incoherence assumptions to avoid "spiky" matrices. These assumptions have been made less stringent in more recent results (Negahban & Wainwright, 2012), which moreover extend the guarantees to approximately low–rank matrices. Such (approximate) low–rank structure is one instance of general structural constraints which are now understood to be necessary for consistent statistical estimation under high–dimensional settings (with very large number of parameters and very few observations). Note that the high–dimensional matrix completion problem is particularly ill–posed, since the measurements are typically both very local (e.g. individual matrix entries), and partial (e.g. covering a decaying fraction of entries of the entire matrix). However, the specific (approximately) low–rank structural constraint imposed in the past work on matrix completion does not capture the rich variety of other qualitatively different structural constraints such as row–sparseness, column–sparseness, or a superposition structure of low–rank plus elementwise sparseness, among others. For instance, in the classical introductory survey on matrix completion (Laurent, 2009), the authors discuss structural constraints of a *contraction matrix*, and a *Euclidean distance matrix*. Thus, the *third question* we ask in this paper is whether we can generalize the recent line of work on low–rank matrix completion to the more general structurally constrained case.

In this paper, we answer all of the three questions above in the affirmative, and provide a vastly unified framework for generalized matrix completion. We address the first two questions by considering a general matrix completion setting wherein observed matrix entries are sampled from any member of a rich family of *natural exponential family distributions*. Note that this family of distributions encompass a wide variety of popular distributions including Gaussian, Poisson, binomial, negative–binomial, Bernoulli, etc. Moreover, the choice of the exponential family distribution can be made depending on the form of the data. For instance, thin–tailed continuous data is typically modeled using the Gaussian distribution; count–data is modeled through an appropriate distribution over integers (Poisson, binomial, etc.), binary data through Bernoulli, categorical–discrete through multinomial, etc. We address the last question by considering general structural constraints upon the underlying matrix, as captured by a general regularization function $\mathcal{R}(.)$. Our general matrix completion setting thus captures heterogeneous noise–channels, for heterogeneous data–types, and heterogeneous structural constraints.

In a key contribution, we propose a simple regularized convex M–estimator for recovering the structurally constrained underlying matrix in this general setting; and moreover provide a unified and novel statistical analysis for our general matrix completion problem. Following a standard approach (Negahban, 2012), we (a) first showed that the negative log–likelihood of the subset of observed entries satisfies a form of Restricted Strong Convexity (RSC) (Definition 4); and (b) under this RSC condition, our proposed M–estimator satisfies strong statistical guarantees. We note that proving these individual components for our general matrix completion problem under general structural constraints required a fairly delicate and novel analysis, particularly the first component of showing the RSC condition, which we believe would be of independent interest. A key corollary of our general framework is matrix completion under sub–Gaussian samples and low–rank constraints, where we show that our theorem recovers results comparable to the existing literature (Candes & Plan, 2010; Keshavan et al., 2010b; Negahban & Wainwright, 2012). Finally, we corroborate our theoretical findings via simulated experiments.

1.1. Notations and Preliminaries

In this subsection we describe the notations and definitions frequently used throughout the paper. Matrices are denoted by capital letters, X, Θ, M, etc. For a matrix M, M_j and $M^{(i)}$ are the j^{th} column and i^{th} row of M respectively, and M_{ij} denotes the $(i,j)^{th}$ entry of M. The *transpose*, *trace*, and *rank* of a matrix M are denoted by $M^\dagger$, $\mathrm{tr}(M)$, and $\mathrm{rk}(M)$, respectively. The inner product between two matrices is given by $\langle X, Y\rangle = \mathrm{tr}(X^\dagger Y) = \sum_{(i,j)} X_{ij}Y_{ij}$.

For a matrix $M \in \mathbb{R}^{m\times n}$ of rank r, with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r$, commonly used matrix norms include the *nuclear norm* $\|M\|_* = \sum_i \sigma_i$, the *spectral norm* $\|M\|_2 = \sigma_1$, the *Frobenius norm* $\|M\|_F = \sum_i \sigma_i^2$, and the *maximum norm* $\|M\|_{\max} = \max_{(i,j)} M_{ij}$.

Given any matrix norm $\|\cdot\|$, the *dual norm*, $\|\cdot\|^*$ is given by $\|X\|^* = \sup_{\|Y\|\leq 1}\langle X, Y\rangle$.

Definition 1 (Natural Exponential Family)**.** A distribution of a random variable X, in a normed vector space is said to belong to the *natural exponential family*, if its probability density function, characterized by the parameter Θ in the dual vector space, is given by:

$$P(X|\Theta) = h(X)\exp\Big(\langle X, \Theta\rangle - G(\Theta)\Big),$$

where $G(\Theta) = \log\int_X h(X)e^{\langle X,\Theta\rangle}\mathrm{dX}$, called the log–partition function, is strictly convex, and analytic.

Definition 2 (Bregman Divergence)**.** Let $\phi : \mathrm{dom}(\phi) \to \mathbb{R}$ be a strictly convex function differentiable in the relative interior of $\mathrm{dom}(\phi)$. The *Bregman divergence* (associated with ϕ) between $x \in \mathrm{dom}(\phi)$ and $y \in \mathrm{ri}(\mathrm{dom}(\phi))$ is defined as:

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle\nabla\phi(y), x - y\rangle.$$

Definition 3 (Subspace compatibility constants)**.** Given a matrix norm $\mathcal{R}(.)$, we define the following maximum and minimum *subspace compatibility* constants of $\mathcal{R}(.)$ w.r.t the subspace $\mathcal{M}$:

$$\Psi(\mathcal{M};\mathcal{R}) = \sup_{\Theta\in\mathcal{M}\setminus\{0\}}\frac{\mathcal{R}(\Theta)}{\|\Theta\|_F}, \quad \Psi_{\min}(\mathcal{R}) = \inf_{\Theta\neq\{0\}}\frac{\mathcal{R}(\Theta)}{\|\Theta\|_F}.$$

Thus, $\forall\Theta \in \mathcal{M}$

$$\Psi_{\min}(\mathcal{R})\|\Theta\|_F \le \mathcal{R}(\Theta) \le \Psi(\mathcal{M},\mathcal{R})\|\Theta\|_F.$$

Definition 4 (Restricted Strong Convexity)**.** A loss function $\mathcal{L}$ is said to satisfy *Restricted Strong Convexity* with respected to a subspace S, if for some $\kappa_{\mathcal{L}} > 0$,

$$\mathcal{L}(\Theta+\Delta) - \mathcal{L}(\Theta) - \langle\nabla\mathcal{L}(\Theta), \Delta\rangle \ge \kappa_{\mathcal{L}}\|\Delta\|_F^2, \forall\Delta \in S.$$

Definition 5 (Sub–Gaussian Distributions)**.** A random variable, X, is said to have a sub–Gaussian distribution with parameter b, if $\forall\ s > 0$, the distribution satisfies $E[e^{sX}] \le e^{s^2b^2/2}$. Further, if X is sub–Gaussian with parameter b, then $E[X] = 0$ and $Var(X) \le b^2$ (refer Vershynin (2010)).

2. Exponential Family Matrix Completion

Denote the underlying target matrix by $\Theta^* \in \mathbb{R}^{m\times n}$. We then assume that individual entries Θ^*_{ij} are observed indirectly via a noisy channel: specifically, via a sample drawn from the corresponding member of *natural exponential family* (see Definition 1):

$$P(X_{ij}|\Theta^*_{ij}) = h(X_{ij})\,\exp\left\{X_{ij}\Theta^*_{ij} - G(\Theta^*_{ij})\right\}, \quad (1)$$

where $G : \mathbb{R} \to \mathbb{R}$, is a strictly convex, and analytic function called the log–partition function.

Consider the random matrix $X \in \mathbb{R}^{m\times n}$, where each entry X_{ij} is drawn independently from the corresponding distribution in (1); it can be seen that:

$$\begin{aligned}P(X|\Theta^*) &= \prod_{ij}\left\{h(X_{ij})\,\exp\left\{X_{ij}\Theta^*_{ij} - G(\Theta^*_{ij})\right\}\right\}\\ &= h(X)\exp\left\{\langle X, \Theta^*\rangle - G(\Theta^*)\right\}, \quad (2)\end{aligned}$$

where we overload the notation to denote $G : \mathbb{R}^{m\times n} \to \mathbb{R}$ as $G(\Theta) = \sum_{ij} G(\Theta_{ij})$, and the base measure $h(X)$ as $h(X) = \prod_{ij} h(x_{ij})$.

Uniformly Sampled Observations: In a "fully observed" setting, we would observe all the entries of the observation matrix $X \in \mathbb{R}^{m\times n}$. However, we consider a partially observed setting, where we observe entries over a subset of indices $\Omega \subset [m] \times [n]$. We assume a uniform sampling model, so that

$$\forall\ (i,j) \in \Omega,\ i \sim \text{uniform}([m]),\ j \sim \text{uniform}([n]). \quad (3)$$

Given, Ω, we define the following matrix $\mathcal{P}_\Omega(X)$:

$$\mathcal{P}_\Omega(X)_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise.}\end{cases}$$

The matrix completion task can then be stated as the estimation of Θ^* from the partially observed matrix $\mathcal{P}_\Omega(X)$, where $X \sim P(X|\Theta^*)$. As noted earlier, this problem is ill–posed in general. However, as we will show, under structural constraints imposed on the parameter matrix Θ^*, we are able to design an M–estimator with a near optimal deviation from Θ^*.

2.1. Applications

Gaussian (fixed σ^2) is typically used to model continuous data, $x \in \mathbb{R}$, such as measurements with additive errors, affinity datasets. Here, $G(\theta) = \frac{1}{2}\sigma^2\theta^2$.
Bernoulli is a popular distribution of choice to model binary data, $x \in \{0, 1\}$, with $G(\theta) = \log(1 + e^\theta)$. Some examples of data suitable for Bernoulli model include social networks, gene protein interactions, etc.
Binomial (fixed N) is used to model number of successes in N trials. Here, $x \in \{0, 1, 2, \ldots, N\}$, and $G(\theta) = N\log(1 + e^\theta)$. Some applications include predicting success/failure rate, survey outcomes, etc.
Poisson is used to model count data $x \in \{0, 1, 2, \ldots\}$, such as arrival times, events per unit time, click–throughs among others. Here, $G(\theta) = e^\theta$.
Exponential is often used to model positive valued continuous data $x \in \mathbb{R}_+$, specially inter arrival times between events. Here, $G(\theta) = -\log(-\theta)$.

2.2. Log–likelihood

Denote the gradient map:

$$g(\Theta) \triangleq \nabla G(\Theta) \in \mathbb{R}^{m\times n},\ \text{where } g(\Theta)_{ij} = \frac{\partial G(\Theta)}{\partial\theta_{ij}}.$$

It can then be verified that the mean and variance of the distribution $P(X|\Theta^*)$ are given by $\mathbb{E}[X] = g(\Theta^*)$, and

$\text{Var}(X) = \nabla^2 G(\Theta^*)$, respectively. The Fenchel conjugate of the log partition function G, is denoted by: $F(X) \triangleq \sup_\Theta \langle X, \Theta \rangle - G(\Theta)$.

A useful consequence of the exponential family is that the negative log–likelihood is convex in the natural parameters Θ^*, and moreover has a bijection with a large class of *Bregman divergences* (Definition 2). The following relationship was first noted by (Forster & Warmuth, 2002), and later established by (Banerjee et al., 2005) [Theorem 4]:

$$-\log P(X|\Theta) \propto B_F(X, g(\Theta)), \ \forall X \in \text{dom(F)}. \quad (4)$$

2.3. Discussion and directions for future work

We consider the standard matrix–completion setting where the distribution of the observation matrix X in (2) corresponds to its entries X_{ij} being drawn independently from the other entries. Further, the probability of observing a specific entry X_{ij}, under uniform sampling is independent of the noise channel or the distribution $P(X_{ij}|\Theta^*_{ij})$. However, in some applications, it might be beneficial to have a sampling scheme involving dependencies; for instance, when a user gives a movie a bad rating, we might want to vary the sampling scheme to sample an entirely different region of the matrix. It would be interesting to extend the analysis of this paper to such a dependent sampling setting.

The form of the observation random matrix distribution in (2), given the individual observations in (1), can be seen to have connotations of multi–task learning: here recovering each individual matrix entry Θ^*_{ij} together with its corresponding noise model forms a single task, and all these tasks can be performed jointly given the shared structural constraint on Θ^*. It would be interesting to generalize this to form a more general statistical framework for *partial* multi-task learning.

We use the general class of exponential family distributions as the underlying probabilistic model capturing the measurement uncertainties. The richness of the class of exponential family distributions has been used in other settings to provide general statistical frameworks. Kakade et al. (2010) provide a generalization of compressed sensing problem to general exponential family distributions. Note however that results from compressed sensing cannot be immediately extended to matrix completion case, since the sampling operator $\mathcal{P}_\Omega$ does not satisfy the typical assumptions (restricted isometry or restricted eigenvalues) made in such settings; see (Candes & Recht, 2009) for additional discussion. There have been extensions of classical probabilistic PCA (Tipping & Bishop, 1999) from Gaussian noise models to exponential family distributions Collins et al. (2001); Mohamed et al. (2008); Gordon (2002). There have also been recent extensions of probabilistic graphical model classes, beyond Gaussian and Ising models, to multivariate extensions of exponential family distributions (Yang et al., 2012; 2013).

More complicated probabilistic models have also been proposed in the context of collaborative filtering (Mnih & Salakhutdinov, 2007; Salakhutdinov & Mnih, 2008), but these typically involve non–convex optimization, and it is difficult to extend the rigorous statistical analyses of the form in this paper (and in the matrix completion literature) to these models.

3. Main Result and Consequences

As noted in the introduction, we consider the matrix completion setting with general structural constraints on the underlying target matrix Θ^*. To formalize the notion of such structural constraints, we follow (Negahban, 2012), and assume that Θ^* satisfies $\Theta^* \in \mathcal{M} \subseteq \overline{\mathcal{M}} \subset R^{m \times n}$, for some subspace $\mathcal{M} \subseteq \overline{\mathcal{M}}$, which contains parameter matrices that are structured similar to the target (the corresponding structural constraints such as low rankness, low rankness+sparsity etc); we also allow the flexibility of working with a superset $\overline{\mathcal{M}}$ of the model subspace that is potentially easier to analyze. Moreover, we use their definition of a decomposable norm regularization function, $\mathcal{R}(.) : \mathbb{R}^{m \times n} \to R_+$, which suitably captures these structural constraints:

A 1. *(Decomposable Norm Regularizer) We assume that $\mathcal{R}(.)$ is a matrix norm, and is decomposable over $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, i.e. if $X \in \mathcal{M}$, $Y \in \overline{\mathcal{M}}^\perp$, then,*

$$\mathcal{R}(X + Y) = \mathcal{R}(X) + \mathcal{R}(Y).$$

We provide some examples of such decomposable regularizers and structural constraint subspaces, and refer to (Negahban, 2012) for more examples and discussion.

Example 1. Low–rank: This is a common structure assumed in numerous matrix estimation problems, particularly those in collaborative filtering, PCA, spectral clustering, etc. The corresponding structural constraint subspaces $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ in this case correspond to a linear span of specific one–rank matrices; we discuss these in further detail in Section 3.2, where we derive a corollary of our general theorem to the specific case of recovery guarantees for low–rank constrained matrix completion. The nuclear norm $\mathcal{R}(\Theta) = \|\Theta\|_* = \sum_k \sigma_k$, has been shown to be *decomposable* with respect to these constraint subspaces (Fazel et al., 2001).

Example 2. Block sparsity: Another important structural constraint for a matrix is block–sparsity, where each row is either all zeros or mostly non–zero, and the number of non–zero rows is small. The structural constraint subspaces in this case correspond to a linear span of specific Frobenius–norm–one matrices that are non–zero in a small subset of the rows (dependent on Θ^*); it has been shown that ℓ_1/ℓ_q

$(q > 1)$ norms (Negahban & Wainwright, 2008; Obozinski et al., 2011) are decomposable with respect to such structural constraint subspaces. Recalling that $\Theta^{(i)}$ is the i^{th} row of Θ, the ℓ_1/ℓ_q norm is defined as:

$$\|\Theta\|_{1,q} = \sum_{i=1}^{m} \|\Theta^{(i)}\|_q = \sum_{i=1}^{m} \Big[\Big(\sum_{j=1}^{n} |\Theta_{ij}|^q \Big)^{1/q} \Big].$$

Example 3. Low–rank plus sparse: This structure is often used to model low–rank matrices which are corrupted by a sparse outlier noise matrix. The structural constraint subspaces corresponding to these consist of the linear span of weighted sum of specific rank–one matrices and sparse matrices with non–zero components on specified positions; and appropriate regularization function decomposable with respect to such structural constraints is the infimum convolution of the weighted nuclear norm with weighted elementwise ℓ_1 norm, $\|M\|_{1,1} = \sum_{ij} |M_{ij}|$ (Candes et al., 2011; Yang & Ravikumar, 2013):

$$\mathcal{R}(\Theta) = \inf\{\lambda_1 \|S\|_{1,1} + \lambda_2 \|L\|_* : \Theta = S + L\}.$$

The second assumption we make is on the curvature of the log–partition function. This is required to establish a form of RSC (Definition 4) for the loss function.

A 2. *The second derivative of the log–partition function $G : \mathbb{R} \to \mathbb{R}$ has atmost an exponential decay, i.e,*

$$\nabla^2 G(u) \geq e^{-\eta|u|}, \ \forall\, u \in \mathbb{R}, \text{ for some } \eta > 0$$

It can be verified that commonly used members of natural exponential family obey this assumption.

Finally, we make an assumption to avoid "spiky" target matrices. As Candes & Recht (2009) show with numerous examples, low–rank and presumably other such structural constraints as above, by themselves are not sufficient for accurate recovery, in part due to the infeasibility of recovering overly "spiky" matrices with very few large entries. Early work (Candes & Plan, 2010; Keshavan et al., 2010a;b), assumed stringent matrix incoherence conditions to preclude such matrices, while more recent work (Davenport et al., 2012; Negahban & Wainwright, 2012), relax these assumptions to restricting the **spikiness ratio**, defined as follows:

$$\alpha_{\text{sp}}(\Theta) = \frac{\sqrt{mn}\|\Theta\|_{\max}}{\|\Theta\|_F}. \tag{5}$$

A 3. *There exists a known $\alpha^* > 0$, such that*

$$\|\Theta^*\|_{\max} = \frac{\alpha_{sp}(\Theta^*)}{\sqrt{mn}} \|\Theta^*\|_F \leq \frac{\alpha^*}{\sqrt{mn}}.$$

3.1. M–estimator for Generalized Matrix Completion

We propose a regularized M–estimate as our candidate parameter matrix $\widehat{\Theta}$. The norm regularizer $\mathcal{R}(.)$ used is a convex surrogate for the structural constraints, and is assumed to satisfy A 1. For a suitable $\lambda > 0$,

$$\begin{aligned}\widehat{\Theta} &= \underset{\|\Theta\|_{\max} \leq \frac{\alpha^*}{\sqrt{mn}}}{\text{argmin}} \frac{mn}{|\Omega|} \Big[\sum_{ij \in \Omega} -\log P(X_{ij}|\Theta_{ij}) \Big] + \lambda \mathcal{R}(\Theta) \\ &= \underset{\|\Theta\|_{\max} \leq \frac{\alpha^*}{\sqrt{mn}}}{\text{argmin}} \frac{mn}{|\Omega|} \Big[\sum_{ij \in \Omega} G(\Theta_{ij}) - X_{ij}\Theta_{ij} \Big] + \lambda \mathcal{R}(\Theta).\end{aligned} \tag{6}$$

The above optimization problem is a convex program, and can be solved by any off–the shelf convex solvers.

3.2. Main Results

Without loss of generality, suppose that $m \leq n$. Let $\mathcal{R}^*(.) = \sup_{\mathcal{R}(X)\leq 1} \langle X, . \rangle$ be the dual norm of the regularizer $\mathcal{R}(.)$. Further, let $\Psi(\overline{\mathcal{M}})$ and $\Psi_{\min}$ be the maximum and minimum *subspace compatibility constants* of $\mathcal{R}$ w.r.t the model subspace $\overline{\mathcal{M}}$ (Definition 3)*. We next define the following quantity:

$$\kappa_{\mathcal{R}}(n, |\Omega|) := \mathbb{E}\Big[\frac{\sqrt{mn}}{|\Omega|} \mathcal{R}^*\Big(\sum_{ij \in \Omega} \epsilon_{ij} e_i e_j^* \Big) \Big],$$

where the expectation is over the random sampling index set Ω, and over the Rademacher sequence $\{\epsilon_{ij} : \forall (i,j) \in \Omega\}$; here $\{e_i \in \mathbb{R}^m\}$, $\{e_j \in \mathbb{R}^n\}$ are the standard basis. This quantity $\kappa_{\mathcal{R}}(n, |\Omega|)$ captures the interaction between the sampling scheme and the structural constraint as captured by the regularizer (specifically its dual $\mathcal{R}^*$). Note that it depends only on n $(n \geq m)$, and on the size $|\Omega|$ of Ω.

Theorem 1. *Let $\widehat{\Theta}$ be the estimate from (6) with $\frac{\lambda}{2} \geq \frac{mn}{|\Omega|} \mathcal{R}^*(\mathcal{P}_\Omega(X - g(\Theta^*))$. Under the assumptions A1-3, if $|\Omega| = \mathbf{\Omega}(\Psi^2(\overline{\mathcal{M}}) n \log n)$†, then given a constant c_0, $\exists$ constants C, C_1, C_2, and K_1, such that, with probability $> 1 - C_1 e^{-C_2 \Psi_{\min} n \log n}$:*

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq C \max\{\alpha^{*2}, 1\} \Psi^2(\overline{\mathcal{M}}) \max\Big\{ \frac{\lambda^2}{\kappa_{\mathcal{L}}^2}, \frac{c_0^2 n \log n}{|\Omega|} \Big\},$$

provided $\kappa_{\mathcal{L}} := e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}} \Big(K_1 - \frac{64}{c_0} \sqrt{\frac{|\Omega| \kappa_{\mathcal{R}}^2(n, |\Omega|)}{n \log n}} \Big) > 0.$

In the above theorem, η and $\alpha^* \geq \alpha_{sp}(\Theta^*)\|\Theta^*\|_F$ are constants from Assumptions 2 and 3, respectively. Our proof uses elements from Negahban (2012), as well as Negahban & Wainwright (2012) where they analyze the case of low–rank structure and additive noise, and establish a form of restricted strong convexity (RSC) for squared loss over subset of matrix entries (closely relates to the special case, when the exponential family distribution assumed in (2) is Gaussian). However, showing such an RSC condition for our general loss function over a subset of structured matrix entries involved some delicate arguments. Further, we provide a much simpler proof that moreover only required a low–spikiness condition rather than a multiplicative spik-

*We suppress the dependence of Ψ and $\Psi_{\min}$ on $\mathcal{R}$ in our notation to avoid notational clutter

†$f(n) = \mathbf{\Omega}(g(n))$ if $f(n) > kg(n)\ \forall n > \hat{n}$.

iness and structural constraint. Moreover, we are able to provide an RSC condition broadly for general structures, and the negative log–likelihood loss associated with general exponential family distribution.

3.3. Corollary

An important special case of the problem is when the parameter matrix Θ^*, is assumed to be of a low–rank $r \ll m$. The most commonly used convex regularizer to enforce low–rank is the nuclear norm. The intuition behind the low–rank assumption on the parameter matrix is as follows: the parameter Θ^*_{ij}, can be thought of as the true measure of affinity between the entities corresponding to row i and column j, respectively; and the data X_{ij}, is a sample from a noisy channel parametrized by Θ_{ij}. It is hypothesized that $\{\Theta^*_{ij}\}$, are obtained from a weighted interaction of a small number of latent variables as, $\Theta^*_{ij} = \sum_{k=1}^{r} \sigma_k u_{ik} v_{jk}$.

Let $\{\mathbf{u_k} \in \mathbb{R}^m\}$ and $\{\mathbf{v_k} \in \mathbb{R}^n\}$, $k \in [r]$ be the left and right singular vectors, respectively of Θ^*. Let the column and row span of Θ^* be $U^* \triangleq \text{col}(\Theta^*) = \text{span}\{\mathbf{u_i}\}$ and $V^* \triangleq \text{row}(\Theta^*) = \text{span}\{\mathbf{v_j}\}$, respectively. Define:

$$\begin{aligned} &\mathcal{M} := \{\Theta : \text{row}(\Theta) \subseteq V^*,\ \text{col}(\Theta) \subseteq U^*\}, \text{ and} \\ &\overline{\mathcal{M}}^{\perp} := \{\Theta : \text{row}(\Theta) \subseteq V^{*\perp},\ \text{col}(\Theta) \subseteq U^{*\perp}\}. \end{aligned} \tag{7}$$

It can be verified that, $\mathcal{M} \neq \overline{\mathcal{M}}$, however, $\mathcal{M} \subset \overline{\mathcal{M}}$.

Corollary 1. *Let Θ^* be a low–rank matrix of rank atmost $r \ll m$. If further, $\forall (i,j)$, $(X_{ij} - g(\Theta^*_{ij}))$ is sub–Gaussian (Definition 5) with parameter b, and $|\Omega| > \mathbf{\Omega}(rn \log n)$, then using $\mathcal{R}(.) = \|.\|_*$ and $\frac{\lambda}{2} := C\sqrt{mn}b\sqrt{\frac{n \log n}{|\Omega|}}$ in (6), w.p. $> 1 - C_1' e^{-C_2' \log n}$,*

$$\frac{1}{mn}\|\widehat{\Theta} - \Theta^*\|_F^2 \leq C' \frac{\max\{\alpha^{*2}, 1\} b^2}{\kappa_{\mathcal{L}}'^2} \left(\frac{rn \log n}{|\Omega|}\right),$$

where $\kappa_{\mathcal{L}}' = K_1' e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}}$.

Remark 1: Note that the above results hold for the minimizer $\widehat{\Theta}$ of the convex program in (6), optimized for any $\alpha^* \geq \alpha_{sp}(\Theta^*)\|\Theta^*\|_F$; in particular it holds with $\alpha^* = \alpha_{sp}(\Theta^*)\|\Theta^*\|_F$, where $1 \leq \alpha_{sp}(\Theta^*) \leq \sqrt{mn}$. While in practice α^* is chosen through cross–validation, the theoretical bound in Corollary 1 can be tightened to the following (if $\|\Theta\|_F \geq 1$):

$$\frac{1}{mn}\frac{\|\widehat{\Theta} - \Theta^*\|_F^2}{\|\Theta^*\|_F^2} \leq C' \frac{\alpha_{\text{sp}}^2(\Theta^*) b^2}{\kappa_{\mathcal{L}}'^2} \left(\frac{rn \log n}{|\Omega|}\right). \tag{8}$$

Similar bound can be obtained for Theorem 1.

Remark 2: The parameter b^2 is a measure of noise per entry in the system; $\forall ij$, $\text{Var}(X_{ij} - g(\Theta^*_{ij})) \leq b^2$.

4. Proof

In this section, we provide key steps in the proofs of the main results (Sec. 3.2-3.3).

4.1. Proof of Theorem 1

The proof of our main theorem involves two key steps:

- We first show that, under assumptions A1-3, *RSC* of the form in Definition 4 holds for the loss function in (6) over a *large subset* of the solution space.
- When the RSC condition holds, the result follows from a few simple calculations; we handle the case where RSC does not hold separately.

Let $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$. We state two results of interest.

Lemma 1. *We define the following subset:*

$$\mathcal{V} = \{\Theta \in \mathbb{R}^{m \times n} : \mathcal{R}(\Theta_{\overline{\mathcal{M}}^{\perp}}) \leq 3\mathcal{R}(\Theta_{\overline{\mathcal{M}}})\},$$

where recall that $\Theta_{\overline{\mathcal{M}}}$ is the projection of Θ onto the subspace $\overline{\mathcal{M}}$. If $\widehat{\Theta}$ is the minimizer of (6), and $\frac{\lambda}{2} \geq \frac{mn}{|\Omega|}\mathcal{R}^*(\mathcal{P}_\Omega(X - g(\Theta^*))$, then $\widehat{\Delta} = \widehat{\Theta} - \Theta^* \in \mathcal{V}$.

The proof follows from Lemma 1 of Negahban (2012).

Lemma 2. *Let $\widehat{\Theta}$ be the minimizer of (6). If $\frac{\lambda}{2} \geq \frac{mn}{|\Omega|}\mathcal{R}^*(\mathcal{P}_\Omega(X - g(\Theta^*))$, then:*

$$\frac{mn}{|\Omega|} \sum_{(i,j)\in\Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \leq \frac{3\lambda\Psi(\overline{\mathcal{M}})}{2}\|\Theta^* - \widehat{\Theta}\|_F$$

The proof is provided in Appendix $A.2$.

Recall the notation $\alpha_{\text{sp}}(\Delta) = \frac{\sqrt{mn}\|\Delta\|_{\max}}{\|\Delta\|_F}$. We now consider two cases, depending on whether the following condition holds for some constant $c_0 > 0$:

$$\alpha_{\text{sp}}(\widehat{\Delta}) \leq \frac{1}{c_0 \Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{n \log n}}. \tag{9}$$

Case 1: Suppose condition in (9) does not hold; so that $\alpha_{\text{sp}}(\widehat{\Delta}) > \frac{1}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{n\log n}}$. From the constraints of the optimization problem (6), we have that $\|\widehat{\Delta}\|_{\max} \leq \|\widehat{\Theta}\|_{\max} + \|\Theta^*\|_{\max} \leq (2\alpha^*/\sqrt{mn})$. Thus,

$$\|\widehat{\Delta}\|_F = \frac{\sqrt{mn}\|\widehat{\Delta}\|_{\max}}{\alpha_{\text{sp}}(\widehat{\Delta})} \leq 2c_0\alpha^* \sqrt{\frac{\Psi^2(\overline{\mathcal{M}}) n \log n}{|\Omega|}}. \tag{10}$$

Case 2: Suppose condition in (9) does hold. Then, the following theorem shows that an RSC condition of the form in Definition 4 holds.

Theorem 2 (Restricted Strong Convexity)**.** *If for a given c_0, $\alpha_{sp}(\widehat{\Delta}) \leq \frac{1}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{n\log n}}$. then, under the assumptions and notations in Theorem 1, w.p. $> 1 - C_1 e^{-C_2 \Psi_{\min} n \log n}$:*

$$\frac{mn}{|\Omega|} \sum_{ij\in\Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \geq \kappa_{\mathcal{L}} \|\widehat{\Delta}\|_F^2,$$

where $\kappa_{\mathcal{L}} = e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}}\left(K_1 - \frac{64}{c_0}\sqrt{\frac{|\Omega|\kappa_{\mathcal{R}}^2(n,|\Omega|)}{n\log n}}\right)$, $K_1 > 0$.

As noted earlier, such an RSC result for the special case of squared loss under low–rank constraints was shown in Negahban & Wainwright (2012). However, proving the RSC

condition for our general setting required a different and more involved proof technique. We prove this theorem in Section 4.3.

Remaining steps of the proof of Theorem 1: Thus, if $\alpha_{\text{sp}}(\widehat{\Delta}) \leq \frac{1}{c_0 \Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{n \log n}}$, and $\kappa_{\mathcal{L}} > 0$, from Theorem 2 and Lemma 2, w.h.p.:

$$\kappa_{\mathcal{L}} \|\widehat{\Delta}\|_F^2 \leq \frac{mn}{|\Omega|} \sum_{ij \in \Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \leq \frac{3\lambda \Psi(\overline{\mathcal{M}})}{2} \|\widehat{\Delta}\|_F \quad (11)$$

From (10) and (11), under assumptions of Theorem 1, we have that for an appropriate constant C, with probability higher than $1 - C_1 e^{-C_2 \Psi_{\min} n \log n}$,

$$\|\widehat{\Delta}\|_F^2 \leq C \max\{\alpha^{*2}, 1\} \Psi^2(\overline{\mathcal{M}}) \max\left\{\frac{\lambda^2}{\kappa_{\mathcal{L}}^2}, \frac{c_0^2 n \log n}{|\Omega|}\right\}.$$

4.2. Proof of Corollary 1

From the definition of $\overline{\mathcal{M}}^{\perp}$ in (7), we have $\overline{\mathcal{M}} = \text{span}\{\mathbf{u_i} x^\dagger, y \mathbf{v_j}^\dagger : x \in \mathbb{R}^n,\ y \in \mathbb{R}^m\}$. Let $P_{U^*} \in \mathbb{R}^{m \times m}$ and $P_{V^*} \in \mathbb{R}^{n \times n}$, be the projection matrices onto the column and row spaces (U^*, V^*) of Θ^*, respectively. We have, $\forall X \in \mathbb{R}^{m \times n}$, $X_{\overline{\mathcal{M}}} = P_{U^*} X + X P_{V^*} - P_{U^*} X P_{V^*}$. Also, $\text{rk}(P_{U^*}) = \text{rk}(P_{V^*}) = \text{rk}(\Theta^*) = r$. Thus, $\forall \Phi \in \overline{\mathcal{M}}$, $\text{rk}(\Phi) \leq 2r$; and hence,

$\Psi(\overline{\mathcal{M}}) = \sup_{\Phi \in \overline{\mathcal{M}} \setminus \{0\}} \frac{\|\Phi\|_*}{\|\Phi\|_F} \leq \sqrt{2r}$. Further, $\Psi_{\min} \geq 1$.

Next, we use the following proposition by Negahban & Wainwright (2012), to bound $\kappa_{\mathcal{R}}(n, |\Omega|)$ in Theorem 1.

Lemma 3 (Lemma 6 of Negahban & Wainwright (2012))**.** *If $\Omega \subset [m] \times [n]$ is sampled using uniform sampling and $|\Omega| > n \log n$, then for a Rademacher sequence $\{\epsilon_{ij}, \forall (i,j) \in \Omega\}$,*

$$\mathbb{E}\Big[\frac{1}{|\Omega|} \| \sum_{ij \in \Omega} \sqrt{mn} \epsilon_{ij} e_i e_j^* \|_2\Big] \leq 10 \sqrt{\frac{n \log n}{|\Omega|}}.$$

Thus, for large enough $c_0 > 640$, using $\kappa_{\mathcal{R}}(n, |\Omega|) = 10\sqrt{\frac{n \log n}{|\Omega|}}$ in Theorem 2, for some $K_1' > 0$ we have:

$$\kappa_{\mathcal{L}}' = e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}} \left(K_1 - \frac{640}{c_0}\right) \geq K_1' e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}}. \quad (12)$$

Finally, to prove the corollary, we derive a bound on $\|\mathcal{P}_\Omega(X - g(\Theta^*))\|_2$ using the Ahlswede–Winter Matrix bound (Appendix A.3). Let $\phi(x) = e^{x^2} - 1$; and let $Y^{(ij)} \triangleq \sqrt{mn}(X_{ij} - g(\Theta^*_{ij})) e_i e_j^*$, such that, $\frac{\sqrt{mn}}{|\Omega|} \|\mathcal{P}_\Omega(X - g(\Theta^*))\|_2 = \| \frac{1}{|\Omega|} \sum_{ij \in \Omega} Y^{(ij)} \|_2$.

From the equivalence of sub-Gaussian definitions in Lemma 5.5 of Vershynin (2010), $\|X_{ij} - g(\Theta^*_{ij})\|_\phi \leq c_0 b$, $\forall ij$. Since, $Y^{(ij)}$ has a single element, $\sqrt{mn}(X_{ij} - g(\Theta^*_{ij}))$, we have, $\|Y^{(ij)}\|_\phi \leq c_0 \sqrt{mn} b$. Further,

$$\begin{aligned}
\mathbb{E}[Y^{(ij)^T} Y^{(ij)}] &= \mathbb{E}[mn(X_{ij} - g(\Theta^*_{ij}))^2 e_j e_j^*] \\
&\overset{(a)}{=} mn \mathbb{E}_{(ij \in \Omega)}[\mathbb{E}_X[(X_{ij} - g(\Theta^*_{ij}))^2] e_j e_j^*] \\
&\overset{(b)}{\leq} mnb^2 \mathbb{E}_{(ij \in \Omega)}[e_j e_j^*] \overset{(c)}{=} mnb^2 \frac{1}{n} I_{n \times n}, \quad (13)
\end{aligned}$$

where (a) follows from Fubini's Theorem, (b) follows as $(X_{ij} - g(\Theta^*_{ij}))$ is sub–Gaussian, and (c) follows from the uniform sampling model. Similarly, $\mathbb{E}[Y^{(ij)} Y^{(ij)^T}] = mnb^2 I_{m \times m}$. Define $\sigma_{ij}^2 := \max\{\mathbb{E}[Y^{(ij)^T} Y^{(ij)}], \mathbb{E}[Y^{(ij)} Y^{(ij)^T}]\}$

In Lemma A.3, using $\sigma_{ij}^2 \leq nb^2$, $\sigma^2 := \sum_{ij \in \Omega} \sigma_{ij}^2 = n|\Omega|b^2$, $M = c_0 \sqrt{mn} b \leq c_0 nb$, and $t = |\Omega|\delta$, we have:

$$P\Big(\Big\| \frac{1}{|\Omega|} \sum_{ij \in \Omega} Y^{(ij)} \Big\|_2 \geq \delta\Big) \leq n^2 \max\Big\{e^{-\frac{\delta^2 |\Omega|}{4nb^2}}, e^{-\frac{\delta |\Omega|}{2 c_0 nb}}\Big\}.$$

Further, for all C, using $\delta = Cb\sqrt{\frac{n \log n}{|\Omega|}}$, there exists a large enough C_1, s.t. if $|\Omega| > C_1 n \log n$, then,

$$P\left(\frac{\sqrt{mn}}{|\Omega|} \|\mathcal{P}_\Omega(X - g(\Theta^*))\|_2 \geq Cb\sqrt{\frac{n \log n}{|\Omega|}}\right) \leq e^{-C_1' \log n}. \quad (14)$$

Using $\Psi_{\min} \geq 1$, $\kappa_{\mathcal{L}} = K_1' e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}}$ (from (12)), and $\frac{\lambda}{2} := C\sqrt{mn} b \sqrt{\frac{n \log n}{|\Omega|}}$ in Theorem 1 leads to the corollary as w.h.p. $\frac{\lambda}{2} = C\sqrt{mn} b \sqrt{\frac{n \log n}{|\Omega|}} \geq \frac{mn}{|\Omega|} \|\mathcal{P}_\Omega(X - g(\Theta^*))\|_2$.

4.3. Proof of Theorem 2

This proof uses symmetrization arguments and contractions (Ledoux & Talagrand (1991) Ch.4&6). We observe that, $\forall\ (ij) \in \Omega$, $\exists v_{ij} \in [0,1]$, s.t.

$$\begin{aligned}
B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) &= G(\widehat{\Theta}_{ij}) - G(\Theta^*_{ij}) - g(\Theta^*_{ij})(\widehat{\Theta}_{ij} - \Theta^*_{ij}) \\
&= \nabla^2 G((1 - v_{ij})\Theta^*_{ij} + v_{ij}\widehat{\Theta}_{ij}) \widehat{\Delta}_{ij}^2 \overset{(a)}{\geq} e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}} \widehat{\Delta}_{ij}^2. \quad (15)
\end{aligned}$$

where (a) holds as $|(1 - v_{ij})\Theta^*_{ij} + v_{ij}\widehat{\Theta}_{ij}| \leq \|\Theta^*\|_{\max} + \|\widehat{\Theta}\|_{\max} \leq \frac{2\alpha^*}{\sqrt{mn}}$, and $\nabla^2 G(u) \geq e^{-\eta|u|}$ (A2).

Lemma 4. *Under Theorem 2, consider the subset,*

$$\mathcal{E} = \Big\{\Delta \in \mathcal{V} : \alpha_{sp}(\Delta) \leq \frac{1}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega|}{n \log n}}, \|\Delta\|_F = 1\Big\}.$$

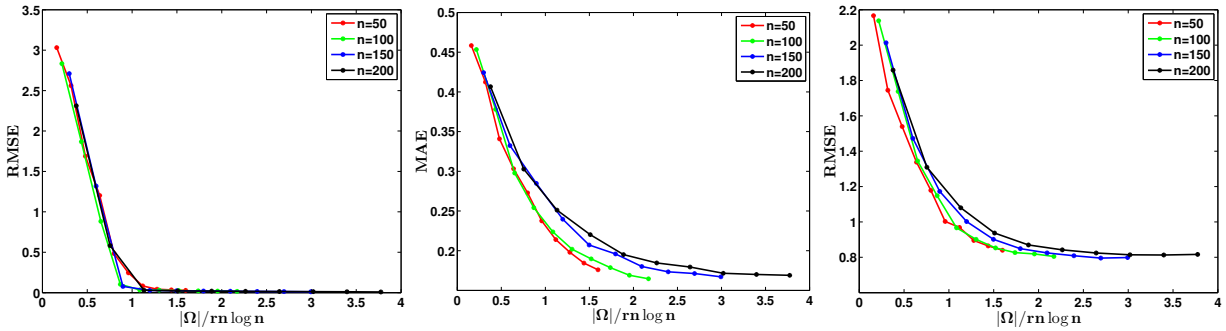

Figure 1. Appropriate error metric between observation matrix X, and the MLE estimate from (6) $\widehat{X}$, plotted against "normalized" sample size, when X is generated from (a) Gaussian, (b) Bernoulli, and (c) binomial distributions

w.p. $> 1 - C_1 e^{-C_2 \Psi_{\min} n \log n}$, $\forall\, \Delta \in \mathcal{E}$:

$$\Big| \frac{mn}{|\Omega|} \sum_{ij \in \Omega} \Delta_{ij}^2 - 1 \Big| \leq \frac{16 \mathcal{R}(\Delta)}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega| \kappa_{\mathcal{R}}^2(n, |\Omega|)}{n \log n}} + k_1' \mathcal{R}(\Delta) \sqrt{\frac{n \log n}{|\Omega|}}.$$

The proof is provided in Appendix $A.1$.

From the assumptions in Thm. 2 and Prop. 1, $\frac{\widehat{\Delta}}{\|\widehat{\Delta}\|_F} \in \mathcal{E}$. Further, as $\widehat{\Delta} \in \mathcal{V}$, $\mathcal{R}(\widehat{\Delta}) = \mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}}) + \mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}) \leq 4\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}}) \leq 4\Psi(\overline{\mathcal{M}}) \|\widehat{\Delta}\|_F$. Using Lemma 4, and (15), and choosing $|\Omega| = c\Psi^2(\overline{\mathcal{M}}) n \log n$, for large enough c, we have $K_1 := 1 - 4k_1' \sqrt{\frac{\Psi^2(\overline{\mathcal{M}}) n \log n}{|\Omega|}} > 0$; thus using $\kappa_{\mathcal{L}} := e^{-\frac{2\eta\alpha^*}{\sqrt{mn}}} \Big(K_1 - \frac{64}{c_0} \sqrt{\frac{|\Omega| \kappa_{\mathcal{R}}^2(n, |\Omega|)}{n \log n}} \Big)$, w.h.p.,

$$\frac{mn}{|\Omega|} \sum_{ij \in \Omega} B_G(\widehat{\Theta}_{ij}, \Theta_{ij}^*) \geq \kappa_{\mathcal{L}} \|\widehat{\Delta}\|_F^2. \tag{16}$$

5. Experiments

We provide simulated experiments to corroborate our theoretical guarantees, focusing specifically on Corollary 1, where we consider the special case where the underlying parameter matrix is low–rank, but the underlying noise model for the matrix elements could be any of the general class of exponential family distributions. We look at three well known members of exponential family suitable for different data–types, namely Gaussian, Bernoulli, and binomial, which are popular choices for modeling continuous valued, binary, and count valued data, respectively.

5.1. Experimental Setup

We create low–rank ground truth parameter matrices, $\Theta^* \in \mathbb{R}^{m \times n}$ of sizes $n \in \{50,\ 100,\ 150,\ 200\}$ (for simplicity we considered square matrices, $m = n$); we set the rank to $r = 2 \log n$. The observation matrices, X, are then sampled from the different members of exponential family distributions parameterized by Θ^*. For each n, we uniformly sample a subset Ω entries of the observation matrix X, and estimate $\widehat{\Theta}$ from (6).

Evaluation:
For each member of the exponential family of distributions considered, we can measure the performance of our M–estimator in parameter space as $\frac{\|\widehat{\Theta} - \Theta^*\|_F^2}{\|\Theta^*\|_F^2}$, or in observation space using an appropriate error metric $\text{err}(\widehat{X}, X)$, where $\widehat{X}$ is the maximum likelihood estimate of the recovered distribution, $\widehat{X} = \text{argmax}_X P(X|\widehat{\Theta})$ (we use RMSE, MAE in our plots). From our corollary, we require $|\Omega| = \mathcal{O}(rn \log n)$ samples for consistent recovery, so we plot the error metric against the the "normalized" sample size, $\frac{|\Omega|}{rn \log n}$. For reasons of space, we only provide results for the error metric in observations space plotted against the the "normalized" sample size. The remainder of the results are provided in Appendix B. It can be seen from the plots that the error decays with increasing sample size, corroborating our consistency results; indeed $|\Omega| > 1.5rn \log n$ samples suffice for the errors to decay to a very small value. Further, the aligning of the curves (for different n) given the "normalized" sample size corroborates the convergence rates as well.

Acknowledgement

The research was funded by NSF Grants IIS-0713142 and IIS-1017614. Pradeep Ravikumar acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894, DMS-1264033.

References

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with bregman divergences. *JMLR*, 6:1705–1749, 2005.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 2010.

Candes, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 2009.

Candes, E. J. and Tao, T. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 2010.

Candes, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Collins, M., Dasgupta, S., and Schapire, R. E. A generalization of principal components analysis to the exponential family. In *NIPS*, pp. 617–624, 2001.

Davenport, M. A., Plan, Y., Berg, E., and Wootters, M. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.

Fazel, M., Hindi, H, and Boyd, S. P. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, pp. 4734–4739, 2001.

Forster, J. and Warmuth, M. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 2002.

Gordon, G. J. Generalizedˆ 2 linearˆ 2 models. In *NIPS*, pp. 577–584, 2002.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.

Kakade, S. M., Shamir, O., Sridharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AISTATS*, JMLR Workshop and Conference Proceedings, pp. 381–388, 2010.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 2010a.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. *JMLR*, 2010b.

Laurent, M. Matrix completion problems. *Encyclopedia of Optimization*, 3:221–229, 2009.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Mnih, A. and Salakhutdinov, R. Probabilistic matrix factorization. In *NIPS*, pp. 1257–1264, 2007.

Mohamed, S., Ghahramani, Z., and Heller, K. A. Bayesian exponential family pca. In *NIPS*, pp. 1089–1096, 2008.

Negahban, S. *Structured Estimation in High-Dimensions*. PhD thesis, EECS Department, University of California, Berkeley, May 2012.

Negahban, S. and Wainwright, M. J. Joint support recovery under high-dimensional scaling: Benefits and perils of l1,-regularization. *NIPS*, 21:1161–1168, 2008.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 98888:1665–1697, 2012.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

Recht, B. A simpler approach to matrix completion. *JMLR*, 12:3413–3430, 2011.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *ICML*, pp. 880–887. ACM, 2008.

Srebro, N., Rennie, J., and Jaakkola, T. S. Maximum-margin matrix factorization. In *NIPS*, pp. 1329–1336, 2004.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 611–622, 1999.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Yang, E. and Ravikumar, P. Dirty statistical models. In *NIPS*, 2013.

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. Graphical models via generalized linear models. In *NIPS*, 2012.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. Conditional random fields via univariate exponential families. In *Advances in Neural Information Processing Systems*, pp. 683–691, 2013.