Abstract

We study the problem of estimating a rank-one matrix from Gaussian observations where different blocks of the matrix are observed under different noise levels. This problem is motivated by applications in clustering and community detection where latent variables can be partitioned into a fixed number of known groups (e.g., users and items) and the blocks of the matrix correspond to different types of pairwise interactions (e.g., user-user, user-item, or item-item interactions). In the setting where the number of blocks is fixed while the number of variables tends to infinity, we prove asymptotically exact formulas for the minimum mean-squared error in estimating both the matrix and the latent variables. These formulas describe the weak recovery thresholds for the problem and reveal invariance properties with respect to certain scalings of the noise variance. We also derive an approximate message passing algorithm and a gradient descent algorithm and show empirically that these algorithms achieve the information-theoretic limits in certain regimes.

1 Introduction

The problem of extracting information from noisy pairwise data appears in applications throughout statistics and machine learning. Over the past two decades, this problem has been studied extensively by focusing on spiked matrix models in which a low-rank signal matrix (the spike) is observed in additive Gaussian noise. The theoretical properties of these models are now well understood in the setting where all of the interactions are of the same type and thus the noise has constant variance. However, much less is known about the more realistic setting in which different types of information are obtained about different types of pairwise interactions.

1.1 Our contribution

This paper provides an asymptotically exact information-theoretic analysis for the general setting in which the underlying objects can be partitioned into to different groups (e.g., users and items) and the strength of the pairwise interactions depends on these groups (e.g., user-user, user-item, or item-item interactions can have different noise levels). Specific contributions include:

- A generalization of the rank-one spiked-matrix model that allows for groupwise heteroskedasticity, i.e., different blocks of the matrix are observed at different noise levels.

- Exact formulas describing information-theoretic limits of this model in the high-dimensional regime where the number of groups is fixed while the number of variables in each group tends to infinity.

- Empirical results demonstrating that the fundamental limits can be attained in some settings using a computationally tractable algorithm based on approximation message passing.
Groupwise spiked matrix model  This paper studies the $K$-group spiked matrix model given by
\[
Y_{kl} = \sqrt{\frac{\lambda_{kl}}{N}} x_k x_\ell^\top + W_{kl}, \quad k, \ell = 1, \ldots, K
\]
where each $x_k$ is an unknown $n_k \times 1$ vector of latent random variables and each $W_{kl}$ is an unknown $n_k \times n_\ell$ noise matrix with independent standard Gaussian entries. The signal-to-noise ratio in each block is parameterized by a non-negative number, $\lambda_{kl}$, and a global scale parameter, $N$, which are assumed to be known. Given the collection of observations $Y = (Y_{kl})$ and the $K \times K$ matrix $\Lambda = (\lambda_{kl})$, the goal is to estimate the latent vectors, $(x_k)$, as well as their rank-one outer products, $(x_k x_\ell^\top)$.

This model can be expressed in matrix form according to
\[
Y = G \circ xx^\top + W,
\]
where $G = (G_{kl})$ is a block constant matrix whose $kl$-th block has value $\sqrt{\lambda_{kl}/N}$, $\circ$ denotes the Hadamard (entry-wise) matrix product, and $x$ is column vector of dimension $n = n_1 + \cdots + n_K$ obtained by stacking $x_1, \ldots, x_K$. We emphasize that $G$ is a known parameter of the model. Hence, if the $\lambda_k$ are strictly positive, then each block of the observations can be rescaled to obtain an equivalent model in which the rank-one signal matrix $xx^\top$ is corrupted by independent additive Gaussian noise whose variance is constant over blocks.

1.2 Comparison with prior work

Low-rank matrix estimation problems have been studied extensively with applications in covariance estimation [25, 26], sparse principal component analysis (PCA) [48], clustering [30, 32], and community detection [16, 1]. The bulk of the theoretical work has focused on one of two basic models: the spiked Wigner model, which corresponds to a single diagonal block of the form $Y_{kk}$ in the groupwise model [1], and the spiked Wishart model, which corresponds to the a single off-diagonal block of the form $Y_{kl}$ in the groupwise model. The statistical limits for these models were studied by Lesieur et al. [31, 32] who used the replica method to derive conjectured formulas for the asymptotic mutual information and MMSE in the setting of independent and identically distributed variables. These formulas were proven rigorously in ensuing work [28, 7, 29, 37] using a variety of interpolation methods from statistical physics. A complementary line of work has focused on the algorithmic limits of recovery associated with spectral methods [5, 11, 18, 43], semidefinite relaxations [2], and approximate message passing (AMP) [20, 17, 41, 32, 38].

More recent work has focused on extending these models in a number of directions, including models with structured priors described by generative models [4, 15, 44], multiview models involving multiple different observations on the same underlying variables [33, 5, 43], and low-rank tensor observation problems [33, 14]. Within this body of work, there has also been significant interest on the existence of computational-to-statistical gaps under various detection and recovery criteria [8, 39].

The idea of integrating different kinds of pairwise information has been studied in the context of clustering problems with high-dimensional covariates [13, 19, 3, 34]. The work most closely related to this paper is the Gaussian version of the contextual stochastic block model [13, 34] which consists of two observation blocks from the groupwise model [1] of the form $(Y_{11}, Y_{12})$, i.e., one obtains pairwise information about the interactions within $x_1$, which are viewed as the target variables of interest, and also the interactions between $x_1$ and $x_2$, where $x_2$ is viewed as an unknown vector of covariates. In this setting, Deshpande et al. [19] characterize the statistical thresholds for detection and recovery and show that these thresholds can be achieved using spectral methods. Ensuing work by Ma and Nandy [34] proves exact formulas for the asymptotic MMSE and also provides a sharp asymptotic analysis of the MSE associated with an AMP algorithm adapted to this setting. In comparison to this work, the contribution of the present paper is that it considers a significantly more general setting that allows for obtains multiple types of observations across the groups.
Figure 1: MSE in estimating the rank-one matrices $uu^\top$ and $vv^\top$ in the two-group model (3) as function of $\alpha$ with $n = 1024$ and $\lambda = 2$. The entries of $x$ are IID standard Gaussian in (a) and IID Rademacher in (b). The asymptotic MMSE (solid line) is given by Theorem 2. The diagonal block MSE is computed by averaging over 64 Monte Carlo trials for each algorithm.

Finally, we also note that our results can be used to address the fundamental limits of the heteroskedastic PCA problem studied by Zhang et al. [46], under the constraint that the noise variance has a block constant structure.

### 1.3 Motivating example

As a motivating example it is useful to consider the special case where there are two groups of equal size and the noise levels are chosen such that the total signal-to-noise level is constant. Specifically, suppose that $u$ and $v$ are independent $n$-dimensional vectors and consider the $2n \times 2n$ model

$$
\begin{pmatrix}
Y_{uu} & Y_{uv} \\
Y_{vu} & Y_{vv}
\end{pmatrix} = \frac{\lambda}{2n} \begin{pmatrix}
\sqrt{1-\alpha} uu^\top & \sqrt{\alpha} uv^\top \\
\sqrt{\alpha} vu^\top & \sqrt{1-\alpha} vv^\top
\end{pmatrix} + W,
\quad x = \begin{pmatrix} u \\ v \end{pmatrix}
$$

(3)

For $\alpha \in \{0, 1/2, 1\}$, this model reduces to the rank-one spiked matrix models studied previously in the literature:

- $\alpha = 0 \iff$ two copies of the $n \times n$ spiked Wigner model with parameter $\lambda$
- $\alpha = 1/2 \iff$ the $2n \times 2n$ spiked Wigner model with parameter $\lambda$
- $\alpha = 1 \iff$ the $n \times n$ spiked Wishart model with parameter $\lambda$

For each of these special cases it is known that if $u$ and $v$ are standard Gaussian vectors then 1) $\lambda^* = 1$ is the critical threshold for weak recovery and 2) there is no statistical-to-computational gap. More specifically, if $\lambda < \lambda^*$, then it is information-theoretically impossible to produce an estimator that has nonzero correlation with the ground truth in the $n \to \infty$ limit. Conversely if $\lambda > \lambda^*$, then estimation with positive correlation is possible and it can be achieved using polynomial-time algorithms based on either spectral methods or approximate message passing.

A natural question of fundamental interest is whether the same threshold behavior also exists in the intermediate regimes where the model is a superposition of these extremes. Using the results in this paper, we are able to provide a positive answer to this question. In particular, we show that if $u$ and $v$ are identically distributed, then the fundamental limits of recovery for the model in (3) are invariant to the parameter $\alpha$. 

3
In the case of Gaussian priors, this means that $\lambda^* = 1$ is the weak recovery threshold for any choice of $\alpha$. Moreover, we demonstrate empirically that for certain prior distributions, these fundamental limits can be achieved using either a gradient based method or approximate message passing; see Figure 4.

An interesting question that we do not resolve is whether this threshold for Gaussian priors can be achieved using spectral methods. The spectral methods considered previously for the integration of different types of information have focused primarily on the problem of estimating $u$ based on only the observations in the block $Y_{uv}$ and $Y_{wv}$. In this setting, Deshpande et al. [19] show that the corresponding weak recovery threshold (which is different from $\lambda^*$) can be achieved by estimating $u$ using the leading eigenvector of the the symmetric matrix $w_{uu}(Y_{uu} + Y_{uv}^T) + w_{uv}(Y_{uv} + Y_{wv}^T)(Y_{uu} + Y_{wv}^T)$ for an appropriate choice of weights $w_{uu}$ and $w_{uv}$. When applied to the setting of model (3), this method is optimal for $\alpha \in \{0, 1\}$ (i.e., the special values of $\alpha$ for which the model reduces to the case of the spiked Wigner or spiked Wishart models) but sub-optimal for positive $\alpha$; see Figure 4. The reason this approach is not optimal is because the estimate of $u$ does not take advantage of the additional information provided in the observations in $Y_{uv}$. Even though $v$ is generated independently of $u$, it is conditionally dependent given the observations and so information about $v$ improves the estimation of $u$.

2 Statement of theoretical results

Our main results are exact formulas for the asymptotic mutual information and MMSE associated with the groupwise model (4). In addition to the pairwise observations, we also allow for side information of the form $w_{uu}Y_{uu}+w_{uv}(Y_{uv}+Y_{wv}^T)(Y_{uu}+Y_{wv}^T)$ for an appropriate choice of weights $w_{uu}$ and $w_{uv}$. When applied to the setting of model (3), this method is optimal for $\alpha \in \{0, 1\}$ (i.e., the special values of $\alpha$ for which the model reduces to the case of the spiked Wigner or spiked Wishart models) but sub-optimal for positive $\alpha$; see Figure 4. The reason this approach is not optimal is because the estimate of $u$ does not take advantage of the additional information provided in the observations in $Y_{uv}$. Even though $v$ is generated independently of $u$, it is conditionally dependent given the observations and so information about $v$ improves the estimation of $u$.

Assumption 1. We consider a sequence of problems, indexed by an integer $N$, with observations from the groupwise spiked model (4) and the side information model (4) such that:

A) The number of groups $K$ is fixed while number of variables in each group $n_k = n_k(N)$ scales as $n_k/N \rightarrow \beta_k \in (0, \infty)$.

B) For each $k = 1, \ldots, K$, the entries of $x_k = (x_{k1}, \ldots, x_{kn_k})$ are drawn independently from a probability measure $P_k$ on $\mathbb{R}$ with second moment equal to one and finite fourth moment.

Rather than working with mutual information directly, we use an equivalent parameterization in terms of the relative entropy between the distribution of the observations and the standard Gaussian measure of the same dimension. For each group $k$, we define the single-letter relative entropy function $D_k: [0, \infty) \rightarrow [0, \infty)$ according to

$$ D_k(\gamma) := D(\mathbb{P}_{\sqrt{\gamma}x_k+w_k} \| \mathbb{P}_{w_k}) $$

where $x_k \sim P_k$ and $w_k \sim \mathcal{N}(0, 1)$ are independent (univariate) random variables and $D(\mathbb{P} \| \mathbb{Q}) = \int \log(\frac{d\mathbb{P}}{d\mathbb{Q}}) d\mathbb{P}$ is the relative entropy (or Kullback-Leibler divergence) between probability measures $\mathbb{P}$ and $\mathbb{Q}$. Associated with the entire collection of observations, we define $D_N: [0, \infty)^K \times [0, \infty)^K \rightarrow [0, \infty)$ to be

$$ D_N(r, A) := \frac{1}{N} D(\mathbb{P}_Y \| \mathbb{P}_W). $$

Our first result shows that the relative entropy $D_N$ converges to a well-defined limit that depends only the parameters $(\beta_k)$ and the single-letter entropy functions associated with each group.
Theorem 1. Under Assumption [4], the relative entropy function $D_N$ converges pointwise to the limit $D$ given by

$$D(r, \Lambda) := \max_q \inf_{\hat{r}} \left\{ \sum_{k=1}^{K} \beta_k D_k(r_k + \hat{r}_k) + \frac{1}{2} q^\top \Lambda q - \frac{1}{2} \hat{r}^\top q \right\}$$

where the maximum is over all $K$-dimensional vectors $q$ such that $0 \leq q \leq \beta$ and the infimum is over all $K$-dimensional non-negative vectors $\hat{r}$.

The variational formula for the limit provides an explicit link between the prior information about the variables, encapsulated by the single-letter relative entropy functions, and the strengths of the different types of interactions, described by the matrix $\Lambda$.

From the point of view of estimation, the significance of Theorem 1 is that changes in the relative entropy with respect to $(r, \Lambda)$ can be related to the MMSE in estimating the underlying variables. Specifically, as a consequence of the I-MMSE relation [21], one finds that $D_N$ is differentiable on the interior of its domain and its partial derivatives satisfy

$$\partial_r D_N(r, \Lambda) = \frac{1}{2N} \mathbb{E} \left[ ||\mathbb{E} [x_k | Y] ||_2^2 \right], \quad \partial_{\lambda_{k\ell}} D_N(r, \Lambda) = \frac{1}{2N^2} \mathbb{E} \left[ ||\mathbb{E} [x_k x_\ell^\top | Y] ||_F^2 \right].$$

In conjunction with Theorem 1, these relations lead to exact formulas for the asymptotic MMSE. The main difference in the asymptotic setting is that there may be a countable number of points at which the limit is non-differentiable. These points correspond to jump-discontinuities in the asymptotic MMSE.

Theorem 2. Consider Assumption [4] and suppose that for a given pair $(r, \Lambda)$ the maximum over $q$ in (7) is attained at a unique point $q^*$. For each group $k$ such that $q^*_k = 0$ or $r_k > 0$ the MMSE satisfies

$$\lim_{N \to \infty} \frac{1}{n_k} \mathbb{E} \left[ ||x_k - \mathbb{E} [x_k | Y] ||_2^2 \right] = \frac{1}{\beta_k} q^*_k,$$

and for each pair $(k, \ell)$ such that $q^*_k q^*_\ell = 0$ or $\lambda_{k\ell} + \lambda_{k\ell} > 0$, the MMSE of $x_k x_\ell^\top$ satisfies

$$\lim_{N \to \infty} \frac{1}{n_k n_\ell} \mathbb{E} \left[ ||x_k x_\ell^\top - \mathbb{E} [x_k x_\ell^\top | Y] ||_F^2 \right] = 1 - \left( \frac{q^*_k}{\beta_k} \right) \left( \frac{q^*_\ell}{\beta_\ell} \right).$$

Furthermore, if the condition $\{q^*_k q^*_\ell = 0\} \cap \{\lambda_{k\ell} + \lambda_{k\ell} > 0\}$ holds for all pairs $(k, \ell)$ then

$$\lim_{N \to \infty} \frac{1}{n^2} \mathbb{E} \left[ ||xx^\top - \mathbb{E} [xx^\top | Y] ||_F^2 \right] = 1 - \left( \sum_{k=1}^{K} q^*_k \right)^2 \left( \sum_{k=1}^{K} \beta_k \right)^2,$$

where $n = n_1 + \cdots + n_k$.

To help interpret these results it is useful to consider some special cases. Observe that if there is no side information $(r = 0)$ and $\Lambda$ is diagonal, then the groupwise model decouples into $K$ independent spiked Wigner models. In this case, the limit can be expressed as

$$\sum_{k=1}^{K} \max_{q_k \in [0, \beta_k]} \left\{ \beta_k D_k(2\lambda_k q_k) - \frac{1}{2} \lambda q_k^2 \right\}.$$

Here, each summand corresponds to the limits obtained previously for the spiked Wigner model [7] with the only difference being a factor of 2 in the signal-to-noise ratio which arises because our model does not assume symmetric noise.
Alternatively, if there is no side information, $\Lambda$ is anti-diagonal, and $K$ is even, then the groupwise model decouples into $K/2$ independent Wishart models. After some straightforward manipulations, the limit can be expressed as

$$
\sum_{k<\ell} \max_{q_k \in \{0, \beta_k\}} \min_{q_\ell \in \{0, \beta_\ell\}} \{ \beta_k D_k (2\bar{\lambda}_k^2 q_k) + \beta_\ell D_\ell (2\bar{\lambda}_\ell^2 q_\ell) - \bar{\lambda}_{k\ell}^2 q_k q_\ell \},
$$

(13)

where $\bar{\lambda}_{k\ell} = (\lambda_{kl} + \lambda_{lk})/2$. Here, each summand corresponds to the limit obtained previously for the spiked Wishart model [37].

The more novel setting occurs when $\Lambda$ has interactions of multiple types. For the two-group model in (3) with $N = 2n$, the limit can be expressed as

$$
\max_{q_1, q_2} \inf_{\tilde{r}_1, \tilde{r}_2} \left\{ \frac{1}{2} D_1(\tilde{r}_1) + \frac{1}{2} D_2(\tilde{r}_2) + \frac{\lambda}{2} [(1 - \alpha)(q_1^2 + q_2^2) + 2\alpha q_1 q_2] - \frac{1}{2} (\tilde{r}_1 q_1 + \tilde{r}_2 q_2) \right\},
$$

(14)

where $D_1$ and $D_2$ are the single-letter relative entropy functions associated with the distributions of $u$ and $v$, respectively.

### 2.1 Proof outline

The key idea underlying the proof of Theorem 1 is that the groupwise model can be expressed as a special case of the matrix tensor product model studied by Reeves [45]. The matrix tensor model was conceived as a generalization of multiview low-rank estimation models in which one obtains multiple observations associated with a collection of $d$-dimensional variables represented by an $n \times d$ matrix $X$. The observations in this model have the form

$$
\hat{Y} = (X \otimes X)B + \hat{W}
$$

(15)

where $X \otimes X$ is the $n^2 \times d^2$ matrix formed by the Kronecker matrix product, $B$ is a known $d^2 \times m$ coupling matrix, and $\hat{W}$ is an $n^2 \times m$ noise matrix with independent Gaussian entries. The main result in [45] is a formula describing the exact limit of the relative entropy in the asymptotic regime where the dimension $d$ is fixed while the number of variables $n$ increases to infinity.

The connection between the groupwise model in (11) and the matrix tensor product model can be understood as follows. First, let the latent vectors $x_1, \ldots, x_K$ be embedded into an $n \times K$ block diagonal matrix of the form $X = \text{diag}(x_1, \ldots, x_K)$. Then, the matrix version of the groupwise model in (2) can be expressed as

$$
Y = \frac{1}{\sqrt{N}} X \Gamma X^T + W, \quad \Gamma := (\sqrt{\lambda_{k\ell}}).
$$

(16)

Next we use vectorization to express this matrix equation in terms of the Kronecker product. For an $m \times n$ matrix $A$, let $\text{vec}(A)$ be the $mn \times 1$ vector obtained by stacking the columns of $A$. We have

$$
\text{vec}(Y) = \frac{1}{\sqrt{N}} (X \otimes X) \text{vec}(\Gamma) + \text{vec}(W).
$$

(17)

Hence, our model corresponds to the matrix tensor product where the dimension is equal to the number of groups and the coupling “matrix” is described by the $k^2 \times 1$ vector $\text{vec}(\Gamma)$.

From here, the remaining step in the proof is to simplify the expression for the limiting relative entropy. A direct application of the results in Reeves [45] leads to a formula that requires optimization over the space of $K \times K$ symmetric positive semidefinite matrices (see [45, Definition 3]). Exploiting the block-diagonal structure of the matrix $X$ arising from the groupwise model, we are able to simplify this expression to obtain the formula given in Theorem 1 in which the optimization is taken over the space of $K$-dimensional nonnegative vectors.

The proof of Theorem 2 follows from Theorem 1 using a similar approach as the proof of [45, Theorem 2]. The full proofs of Theorems 1 and 2 are provided in Appendix A.
3 Methods

To simplify the presentation of the algorithmic results it is convenient to apply a preprocessing step that symmetrizes the observations. The symmetrized observations and parameters are defined according to

\[
Y_{k\ell}^{\text{sym}} = \sqrt{\frac{\lambda_{k\ell}}{\lambda_{k\ell} + \lambda_{\ell k}}} Y_{k\ell} + \sqrt{\frac{\lambda_{\ell k}}{\lambda_{k\ell} + \lambda_{\ell k}}} Y_{\ell k}^\top, \quad \lambda_{k\ell}^{\text{sym}} = \lambda_{k\ell} + \lambda_{\ell k},
\]

where we use the convention \( Y_{k\ell}^{\text{sym}} = \sqrt{\frac{1}{2}} Y_{k\ell} + \frac{1}{2} Y_{\ell k}^\top \) in the case \( \lambda_{k\ell}^{\text{sym}} = 0 \). By the orthogonal invariance of the standard Gaussian distribution, it can be verified that \( Y_{k\ell}^{\text{sym}} \) are sufficient statistics for estimation of \( \mathbf{x}_k \). Furthermore, the distribution of \( Y_{k\ell}^{\text{sym}} \) can be described by the symmetric \( K \)-group model:

\[
Y_{k\ell}^{\text{sym}} = \sqrt{\frac{\lambda_{k\ell}^{\text{sym}}}{N}} \mathbf{x}_k \mathbf{x}_\ell^\top + W_{k\ell}^{\text{sym}}, \quad 1 \leq k \leq \ell \leq K
\]

where \( W_{k\ell}^{\text{sym}} : 1 \leq k \leq \ell \leq K \) are independent Gaussian noise matrices such that each off-diagonal block \( (k \neq \ell) \) has independent standard Gaussian entries and each diagonal block \( (k = \ell) \) is drawn from the Gaussian orthogonal ensemble, i.e., it a symmetric matrix with independent standard Gaussian entries above the diagonal and independent Gaussian entries of mean zero and variance two on the diagonal. The symmetric groupwise model can be written in matrix form according to

\[
Y^{\text{sym}} = G^{\text{sym}} \circ \mathbf{x} \mathbf{x}^\top + W^{\text{sym}},
\]

where \( \mathbf{x} \) is the \( n \times 1 \) vector obtained from stacking the \( \mathbf{x}_k \), \( W^{\text{sym}} \) is an \( n \times n \) symmetric matrix drawn from the Gaussian orthogonal ensemble, and \( G^{\text{sym}} \) is an \( n \times n \) block-constant matrix with entries in the \( kl \)-th block given by \( N^{-1/2} (\lambda_{kl}^{\text{sym}})^{1/2} = N^{-1/2} (\lambda_{k\ell} + \lambda_{\ell k})^{1/2} \).

Remark 1. Because the asymmetric model \([1]\) can be mapped to the symmetric model \([19]\), we could have used the symmetric form as the starting point for our analysis. The reason we present the asymmetric model is that it is more natural for certain problems settings where there is a difference between interactions from group \( k \) to \( \ell \) and the interactions from \( \ell \) to \( k \).

3.1 Spectral methods

Spectral methods have been studied extensively both in the context of single-type interactions (i.e., the spiked Wigner and spiked Wishart models) as well as covariate assisted models and multiview models. In the case of single-type interactions, the spectral methods are optimal when the factors have an orthogonally invariant distribution. We consider two baseline methods based on principal component analysis (PCA):

- **Joint PCA**: Let \( Y^* \) be the symmetric \( n \times n \) data matrix obtained by zeroing out the blocks in \( Y^{\text{sym}} \) with \( \lambda_{kl}^{\text{sym}} = 0 \). The estimate of each \( \mathbf{x}_k \) is chosen to be proportional to the corresponding sub-vector in the leading eigenvector of \( Y^* \).

- **Weighted PCA**: Given nonnegative weights \( w_{kl} \) the estimate of each \( \mathbf{x}_k \) is chosen to be proportional to the leading eigenvector of the the \( n_k \times n_k \) matrix

\[
w_{kk} Y_{kk}^{\text{sym}} + \sum_{\ell \neq k} w_{k\ell} Y_{k\ell}^{\text{sym}} (Y_{k\ell}^{\text{sym}})^\top.
\]

In the case of the two-group model, this method is equivalent to the one proposed by Deshpande et al. \([19]\), Theorem 6), who provide an explicit specification for the weights \( w_{12} \) and \( w_{21} \) as a function of the noise level in each block.
3.2 Gradient descent

Gradient descent algorithms have been studied in many contexts and provide a straightforward optimization procedure. Recent work (e.g., [17, 12]) has shown that, under reasonable conditions, low-rank matrix recovery has an objective without any spurious local minima, even though it is a non-convex problem. A gradient method was also studied in the spiked matrix model under generative priors [15].

Under the symmetric model (20), the negative log likelihood function can be expressed as

$$x \mapsto \frac{1}{4} \|Y_{sym} - G_{sym} \circ x x^\top\|_F^2.$$  (22)

The maximum likelihood estimate is approximated by performing gradient descent on (22). Let \(x^0\) be an initial estimate of \(x\) and let \(\gamma_0, \gamma_1, \ldots\) be a sequence of positive numbers. For \(t = 0, 1, 2, \ldots\) the gradient descent updates are given by

$$x^{t+1} = x^t + \gamma_t \left[ (G_{sym} \circ Y_{sym}) x^t - \left( (G_{sym})^{\circ 2} (x^t)^2 \right) \circ x^t \right]$$  (23)

where \((A)^{\circ 2}\) denotes the entry-wise square for an arbitrary matrix or vector, \(A\). The updates can be expressed equivalently at the group level according to

$$x_{k+1} = x_k^t + \gamma_t \left[ \sum_{\ell=1}^K \frac{\lambda_{\ell k}^{sym}}{N} Y_{k\ell} x_{\ell}^t - \left( \frac{1}{N} \sum_{\ell=1}^K \lambda_{\ell k}^{sym} \|x_{\ell}^t\|^2 \right) \circ x_k^t \right], \quad k = 1, \ldots, K.$$  (24)

3.3 Approximate message passing

A variety of AMP algorithms have been proposed in the context of low-rank matrix estimation problems [20, 17, 11, 32, 38, 34]. This section gives a version of AMP adapted to the symmetric form of the groupwise model (19). The derivation follows the general outline given by Lesieur et al. [32] and is described in full detail in Appendix B.

For \(k = 1, \ldots, K\), define the function

$$\eta_k(a, b) = \int x_k \exp \left\{ ax_k - \frac{1}{2} b x_k^2 \right\} dP_k(x_k),$$

$$\int \exp \left\{ a x_k - \frac{1}{2} b (x_k)^2 \right\} dP_k(x_k),$$  (25)

and let \(\eta'(a, b)\) denote the derivative with respect to the first argument. Also, for each \(k = 1, \ldots, K\), let \(m_k^0\) be an initial estimate of \(x_k\). For \(t = 0, 1, 2, \ldots\) the AMP updates are given by

$$a_k^t = \sum_{\ell=1}^K \frac{\lambda_{k\ell}^{sym}}{N} Y_{k\ell} m_{\ell}^t - \left( \frac{1}{N} \sum_{\ell=1}^K \lambda_{k\ell}^{sym} (Y_{k\ell}^{sym})^{\circ 2} v_{\ell}^t \right) \circ m_{k-1}^{t-1}$$  (26a)

$$b_k^t = \frac{1}{N} \sum_{\ell=1}^K \left( \lambda_{k\ell}^{sym} (m_{\ell}^t)^{\circ 2} + \lambda_{k\ell}^{sym} (1_{n_k \times n_k} - (Y_{k\ell}^{sym})^{\circ 2}) v_{\ell}^t \right)$$  (26b)

$$m_k^{t+1} = \eta_k \left( a_k^t, b_k^t \right)$$  (26c)

$$v_{k}^{t+1} = \eta_k \left( a_k^t, b_k^t \right),$$  (26d)

where \(m_{k-1}^{-1}\) and \(v_k^0\) are initialized to the all zeros vector, the functions \(\eta_k\) and \(\eta'_k\) are applied entrywise to vector-valued inputs, and all of the terms for \(k = 1, \ldots, K\) are updated at each iteration. After \(T\) iterations, the approximation of the conditional mean \(x_k\) is given by \(m_k^T\) and the approximation to the conditional variance is given by \(v_k^T\). Furthermore, the vector \(a_k^T\) can be viewed as an unbiased estimate of \(x_k\). See Appendix B for more details.

4 Numerical results

This section provides numerical results comparing the formulas for the asymptotic MMSE in Theorem \(\text{2}\) and the empirical performance of the methods described in Section \(\text{3}\).
Evaluation of MSE We focus on the two-group model in (3) and performance is assessed in terms of the MSEs in estimating the rank-one matrices \( \mathbf{u}\mathbf{u}^\top \) and \( \mathbf{v}\mathbf{v}^\top \) appearing on the diagonal blocks of \( \mathbf{x}\mathbf{x}^\top \). In all cases, the estimates of the matrices are obtained from the estimates \( \hat{\mathbf{u}} \) and \( \hat{\mathbf{v}} \) of the corresponding vectors and the MSE can be expressed as

\[
\text{mse}_{\mathbf{u}\mathbf{u}} = \frac{1}{n^2} \mathbb{E} \left[ \| \mathbf{u}\mathbf{u}^\top - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top \|_F^2 \right], \quad \text{mse}_{\mathbf{v}\mathbf{v}} = \frac{1}{n^2} \mathbb{E} \left[ \| \mathbf{v}\mathbf{v}^\top - \hat{\mathbf{v}}\hat{\mathbf{v}}^\top \|_F^2 \right]
\]  

(27)

For the gradient descent method and AMP, these expectation are approximated directly via Monte Carlo trials. Using this identity, we obtain an unbiased estimate of the desired MSE by approximating the second expectation on the right-hand side using Monte Carlo trials. Implementation details For the empirical results, the dimension of each block is \( n = 512 \) and the MSE is estimated via 32 Monte Carlo trials. The gradient descent method is initialized using estimates obtained from joint PCA, and AMP is initialized with each \( \mathbf{m}_k \) initialized with IID \( \mathcal{N}(0, 10^{-6}) \) entries. For the weighted PCA method, the weights are chosen via a grid search optimization procedure that minimizes the empirical loss. Note that in doing so, the weights are adapted to the ground truth values of \( \mathbf{u} \) and \( \mathbf{v} \), and thus the resulting MSE should be viewed as a lower bound on the performance of the weighted PCA method for any data-driven selection of the weights. Additional details regarding the numerical approximation of the MMSE formulas are given in Appendix C.

Gaussian priors The case where both \( \mathbf{u} \) and \( \mathbf{v} \) have independent standard Gaussian entries is shown in Figure 1a and Figure 2. In this setting, the asymptotic MMSE is constant for all \( 0 \leq \alpha \leq 1 \) and the weak detection threshold (i.e., the smallest value of \( \lambda \) such that the asymptotic MMSE is strictly less than the MMSE obtained without any observations) occurs at the critical threshold \( \lambda^* = 1 \). The empirical performance of both gradient descent and AMP shows excellent agreement with the formulas for the asymptotic MMSE. By contrast, the spectral methods are sensitive to the choice of the parameter \( \alpha \). More specifically, the spectral methods perform well at the special values of \( \alpha \) for which the two-group model reduces to one of the previously studied spiked matrix models. (These values are given by \( \alpha = 1/2 \) for joint PCA and \( \alpha \in \{0,1\} \) for weighted PCA.) However, as \( \alpha \) deviates from these special values the performance degrades sharply, indicating that these spectral methods are suboptimal in general.

Non-Gaussian priors A comparison of all algorithms considered in Section 3 is given in Figure 1b when both \( \mathbf{u} \) and \( \mathbf{v} \) have non-Gaussian priors. Specifically, \( \mathbf{u} \) and \( \mathbf{v} \) are both distributed IID according to a Rademacher distribution (i.e., each variable is \( \pm 1 \) with equal probability). We observe that only the AMP algorithm approaches the MMSE (which is invariant to \( \alpha \) in this case) and the other methods (gradient descent, joint PCA, and weighted PCA) perform similarly to the Gaussian priors case, which can be seen in Figure 1a. Two additional examples involving non-Gaussian priors are shown in Figure 3. In the first example, both \( \mathbf{u} \) and \( \mathbf{v} \) have independent Rademacher entries and in the second example, \( \mathbf{u} \) has independent Bernoulli(0.1) entries (shifted and scaled to mean zero and variance one) and \( \mathbf{v} \) has independent standard Gaussian entries. In both cases, empirical results are shown only for AMP. The asymptotic MME is invariant.
Figure 2: MSE in estimating the rank-one matrices $uu^\top$ and $vv^\top$ in the two-group model (3) as function of $\lambda$ for various $\alpha$ with $n = 512$ and IID standard Gaussian priors. The asymptotic MMSE (solid line) is given by Theorem 2. The subplots show the diagonal block MSE of joint PCA (upper left), weighted PCA (upper right), gradient descent (lower left), and AMP (lower right).

to $\alpha$ in the first example, where $u$ and $v$ have the same distribution, but depends on $\alpha$ in the second example, where $u$ and $v$ have different distributions. Interestingly, the MMSE curves in the second example are ordered for both the MMSE of $uu^\top$ and $vv^\top$, however, the order is reversed between the two. In particular, smaller values of $\alpha$ result in better performance (lower MMSE) in estimating $uu^\top$, but have worse performance (higher MMSE) in estimating $vv^\top$.

5 Conclusion

This paper considers a block heterogeneous variation of the extensively studied low-rank matrix estimation problem. The information-theoretic limits for this model are derived by embedding it within the matrix tensor-product model studied by Reeves [45]. We derive single-letter formulas for the MMSE that depend on the noise level in each block and the prior distributions of the underlying variables. Complementary to the fundamental limits, we also consider the empirical performance of methods based on spectral decomposition, gradient decent, and AMP.

The assumptions used in our theoretical results can be relaxed to a certain extent. For example, following
Figure 3: MSE in estimating the rank-one matrices $uu^\top$ and $vv^\top$ in the two-group model (3) as function of $\lambda$ for various $\alpha$ with $n = 512$. The asymptotic MMSE (solid line) is given by Theorem 2. The empirical MSE corresponds to AMP. In the top subplot, both $u$ and $v$ are Radamacher. In the bottom subplots, $u$ has a Bernoull(i(0.1)) prior (scaled and shifted to zero mean and unit variance) and $v$ is standard Gaussian. In the lower left subplot, we are plotting the MMSE and MSE in estimating $uu^\top$ and in the lower right subplot, we are plotting the MMSE and MSE in estimating $vv^\top$. 
the approach in [45], the number of groups \( K \) can be allowed to grow at the rate \( N^\alpha \) for sufficiently small positive number \( \alpha \) and the assumption that the entries are identically distributed can also be relaxed provided that per-variable relative entropy function has a well-defined limit.

An interesting open question for future work is whether a properly defined spectral method can achieve the weak recovery threshold in the two-group model [3] for all values of \( \alpha \).

References

[1] Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86.

[2] Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, 37(5):2877–2921.

[3] Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. (2020). Inference for multiple heterogeneous networks with a common invariant subspace.

[4] Aubin, B., Loureiro, B., Maillard, A., Krzakala, F., and Zdeborová, L. (2021). The spiked matrix model with generative priors. *IEEE Transactions on Information Theory*, 67(2):1156–1181.

[5] Baik, J., Arous, G. B., and Pêché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33:1643–1697.

[6] Banks, J., Moore, C., Verzelen, N., Vershynin, R., and Xu, J. (2018). Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894.

[7] Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T., and Zdeborová, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, pages 424–432, Barcelona, Spain.

[8] Barbier, J., Macris, N., and Rush, C. (2020). All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 14915–14926.

[9] Barbier, J. and Reeves, G. (2020). Information-theoretic limits of a multiview low-rank symmetric spiked matrix model. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. [Online]. Available [https://arxiv.org/abs/2005.08017](https://arxiv.org/abs/2005.08017).

[10] Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

[11] Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135.

[12] Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29.

[13] Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377.

[14] Chen, H.-B., Mourrat, J.-C., and Xia, J. (2021). Statistical inference of finite-rank tensors.

[15] Cocola, J., Hand, P., and Voroninski, V. (2020). Nonasymptotic guarantees for spiked matrix recovery with generative priors. In *Advances in Neural Information Processing Systems*, volume 33, pages 15185–15197.
[16] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6).

[17] Deshpande, Y. and Montanari, A. (2014a). Information-theoretically optimal sparse PCA. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201, Honolulu, HI.

[18] Deshpande, Y. and Montanari, A. (2014b). Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–432.

[19] Deshpande, Y., Montanari, A., Mossel, E., and Sen, S. (2018). Contextual stochastic block models. In *NeurIPS*.

[20] Fletcher, A. K. and and, S. R. (2018). Iterative reconstruction of rank-one matrices in noise. *Information and Inference*, 7(3):531–562.

[21] Guo, D., Shamai, S., and Verdú, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282.

[22] Guo, D. and Verdú, S. (2005). Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010.

[23] Guo, D., Wu, Y., Shamai, S., and Verdú, S. (2011). Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4):2371–2385.

[24] Hiriart-Urruty, J.-B. and Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms I*. Springer-Verlag.

[25] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.

[26] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–703.

[27] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

[28] Krzakala, F., Xu, J., and Zdeborová, L. (2016). Mutual information in rank-one matrix estimation. In *Proceedings of the IEEE Information Theory Workshop (ITW)*.

[29] Lelarge, M. and Miolane, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929.

[30] Lesieur, T., de Bacco, C., Banks, J., Krzakala, F., Moore, C., and Zdeborová, L. (2016). Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, pages 601–608, Monticello, IL.

[31] Lesieur, T., Krzakala, F., and Zdeborová, L. (2015). Phase transitions in sparse PCA. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1635–1639, Hong Kong.

[32] Lesieur, T., Krzakala, F., and Zdeborová, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7).

[33] Luneau, C., Barbier, J., and Macris, N. (2020). Mutual information for low-rank even-order symmetric tensor estimation. *Information and Inference: A Journal of the IMA*.

[34] Ma, Z. and Nandy, S. (2021). Community detection with contextual multilayer networkscommunity detection with contextual multilayer networks. [Online]. Available https://arxiv.org/abs/2104.02960v1.
[35] Mayya, V. and Reeves, G. (2019). Mutual information in community detection with covariate information and correlated networks. In Proceedings of the Allerton Conference on Communication, Control, and Computing, pages 602–607, Monticello, IL.

[36] Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. Econometrica, 70(2):583–601.

[37] Miozane, L. (2017). Fundamental limits of low-rank matrix estimation: the non-symmetric case. [Online]. Available https://arxiv.org/abs/1702.00473.

[38] Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. The Annals of Statistics, 49(1):321–345.

[39] Niles-Weed, J. and Zadik, I. (2020). The all-or-nothing phenomenon in sparse tensor pca. In Advances in Neural Information Processing Systems, volume 33, pages 17674–17684.

[40] Pandit, P., Sahraee Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. (2020). Matrix inference and estimation in multi-layer models. In Advances in Neural Information Processing Systems, volume 33, pages 22456–22467.

[41] Parker, J. T., Schniter, P., and Cevher, V. (2014). Bilinear generalized approximate message passing—part I: Derivation. IEEE Transactions on Signal Processing, 62(22):5839–5853.

[42] Pearl, J. (1998). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[43] Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. Annals of Statistics, 46(5):2416–2451.

[44] Polyanskiy, Y. and Wu, Y. (2016). Wasserstein continuity of entropy and outer bounds for interference channels. IEEE Transactions on Information Theory, 62(7):3992–4002.

[45] Reeves, G. (2020). Information-theoretic limits for the matrix tensor product. IEEE Journal on Selected Areas in Information Theory, 1(3):777–798.

[46] Zhang, A. R., Cai, T. T., and Wu, Y. (2021). Heteroskedastic pca: Algorithm, optimality, and applications. [Online]. Available https://arxiv.org/abs/1810.08316v3.

[47] Zhang, J. and Zhang, R. (2020). How many samples is a good initial point worth in low-rank matrix recovery? In Advances in Neural Information Processing Systems, volume 33, pages 12583–12592.

[48] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2):265–286.

A Proofs of main results

As described in Section 2.1 the main idea in the proof is to show that the groupwise model in (1) can be expressed as a special case of the matrix tensor product model studied in [45]. In the following, we first summarize the basic definition and results from [45] and then provide the proofs for Theorems 1 and 2.
A.1 The matrix tensor product model

Let $X$ be an $n \times d$ random matrix. Following [45, Equation (9)], the matrix tensor product model associated with a positive scale factor $N$, $d \times d$, symmetric positive semidefinite matrix $R$, and $d^2 \times d^2$ symmetric positive semidefinite matrix $S$ is defined by

$$ Y_{MTP}^{R,S} = \begin{cases} XR^{1/2} + W \\ \frac{1}{\sqrt{N}}(X \otimes X)S^{1/2} + W' \end{cases} $$

(29)

where $\otimes$ denotes the Kronecker matrix product and $W$ and $W'$ are independent matrices with independent standard Gaussian entries. The only difference between this definition and the one in [45] is that the scale factor $N$ maybe different from the number of rows $n$. The relative entropy function associated with this model is defined by

$$ D_{MTP}^N(R, S) := \frac{1}{N} D(P_{Y_{MTP}^N, 0} \parallel P_{Y_{MTP}^0, 0}), $$

(30)

where $P_{Y_{MTP}^0, 0}$ is the standard Gaussian measure of the same dimension as the observations. The approximation formula (see [45, Equation (42)]) is defined as

$$ \hat{D}_{MTP}^N(R, S) := \max_Q \inf_{\tilde{R} \in \mathbb{S}^d_{++}} \left\{ D_N^{MTP}(R + \tilde{R}, 0) + \frac{1}{2} \text{tr}(S(Q \otimes Q)) - \frac{1}{2} \text{tr}(\tilde{R}Q) \right\}. $$

(31)

where the maximum is over $d \times d$ symmetric positive semidefinite matrices satisfying $Q \preceq \frac{1}{N} \mathbb{E}[X^\top X]$.

The main results in [45] are bounds on the difference between the relative entropy and the approximation formula. For the purposes of this paper, we are interested in the asymptotic setting where the number of rows increases while the number of columns is fixed.

**Theorem 3** ([45]). Consider a sequence of problems, indexed by integer $N$, with observations from the matrix tensor product model (29) such that the number of columns $d$ is fixed while the number of rows scales as $n/N \to \beta \in (0, \infty)$. If the rows of $X$ are independent with $\mathbb{E}[|X_{ij}|^4] \leq C$ for some constant $C$, then, $|D_{MTP}^N - \hat{D}_{MTP}^N|$ converges to zero pointwise as $N \to \infty$.

**Proof.** This result follows along similar lines as [45, Corollary 1]. If the entries of $X$ are bounded uniformly, then the result follows directly from [45, Theorem 1]. The relaxation from bounded entries to bounded forth moment follows from the continuity of relative entropy with respect to the quadratic Wasserstein distance [44] and the fact that a finite fourth moment bound on the entries of $X$ implies a finite second moment bound on the entries of $X \otimes X$. \qed

A.2 Proof of Theorem 1

Let $X = \text{diag}(x_1, \ldots, x_K)$ be the $n \times K$ block-diagonal matrix whose $k$-th block is given by $x_k$ and let $\Gamma$ be the $K \times K$ matrix corresponding to the entrywise positive square root of $\Lambda$. Using this notation, the groupwise model (1) can be expressed in matrix form according to

$$ Y = \frac{1}{\sqrt{N}} X \Gamma X^\top + W. $$

(32)

By vectorization, this model can be written equivalently as

$$ \text{vec}(Y) = \frac{1}{\sqrt{N}} (X \otimes X) \text{vec}(\Gamma) + \text{vec}(W). $$

(33)

This is the matrix tensor product model given in [45, Definition 1] with $K^2 \times 1$ coupling matrix $\text{vec}(\Gamma)$. As noted in [45, Proposition 1], the fundamental limits for this model are the same as for the symmetric
form of the matrix tensor product model (see Definition 2) characterized by $K^2 \times K^2$ coupling matrix $\text{vec}(\Gamma)\text{vec}(\Gamma)$. Recalling that the relative entropy function $D_N(r, \Lambda)$ in (3) is defined with respect to observations from the groupwise model (1) and the side-information model (2) we can write

$$D_N(r, \Lambda) = D_N^{\text{MTP}}(\text{diag}(r), \text{vec}(\Gamma)\text{vec}(\Gamma)^\top)$$

(34)

where $D_N^{\text{MTP}}(R, S)$ is defined in (30).

From here, a direct application of Theorem 3 shows that the limiting behavior of $D_N$ can be expressed in terms of the approximation formula for the matrix tensor product in (31). To simplify the analysis, we use the following result, which leverages the block-diagonal structure in $X$.

**Lemma 4.** Suppose that $X = \text{diag}(x_1, \ldots, x_K)$ is a block-diagonal matrix where each $x_k$ is a column vector. Then, the relative entropy function in the matrix tensor product model depends only on the diagonal entries of the matrices $(R, S)$. Specifically,

$$D_N^{\text{MTP}}(R, S) = D_N^{\text{MTP}}(\text{diag}(r), \text{diag}(s)),$$

(35)

where $r$ and $s$ are vectors containing the diagonal entries in $R$ and $S$, respectively.

**Proof.** Consider the observation model $A u + w$ where $u$ is a $p \times 1$ random vector, $A = [a_1, \ldots, a_K]$ is an $m \times p$ deterministic matrix, and $w$ is an independent $m \times 1$ vector with standard Gaussian entries. Suppose that the entries of $u$ are deterministically zero except in the $k$-th location and let $Q_k$ be an $m \times m$ orthogonal matrix (chosen independently of $u$ and $w$) such that the $k$-th row of $Q_k$ is given by $a_k/\|a_k\|$. Then, we can write

$$D(Au + w \parallel w) = D(a_k u_k + w \parallel w)$$

(36)

$$= D(Q_k a_k u_k + Q_k w \parallel Q_k w)$$

(37)

$$= D((\text{diag}([\|a_1\|, \ldots, \|a_K\|])) u + Q_k w \parallel Q_k w)$$

(38)

where the second line follows from the invariance of relative entropy to one-to-one transformation and the third line follows from first noting that $Q_k a_k = \|a_k\| e_k$, where $e_k$ is the $k$-th standard basis vector, and then recalling that only the $k$-th entry in $u$ is nonzero. By the orthogonal invariance of the standard Gaussian distribution $Q_k w$ is a standard Gaussian vector that is independent of $u$. Now suppose that $T$ is a $p \times p$ symmetric positive semidefinite matrix with diagonal given by $t$. Setting $A = T^{1/2}$ we find that $(\|a_1\|^2, \ldots, \|a_K\|^2) = t$ and so the arguments given above imply that

$$D(T^{1/2} u + w \parallel w) = D(\text{diag}(t)^{1/2} u + w \parallel w).$$

(39)

By the chain rule for relative entropy, this result extends to the setting where $U$ is an $n \times p$ random matrix with the property that each row has at most one entry that is not deterministically zero, i.e., we can write

$$D(U T^{1/2} + W \parallel W) = D(U \text{diag}(t)^{1/2} + W \parallel W).$$

(40)

where $W$ is an independent $n \times p$ matrix with independent standard Gaussian entries. Finally, the stated result for the matrix tensor product model follows from noting that both $X$ and $X \otimes X$ have the property that at most one entry in each row is not deterministically zero.

In view of Lemma 4 the relation in (34) can simplified as

$$D_N(r, \Lambda) = D_N^{\text{MTP}}(\text{diag}(r), \text{diag}(\text{vec}(\Lambda))).$$

(41)

Moreover, if $\tilde{R}$ is a $K \times K$ symmetric positive semidefinite matrix with diagonal given by $\tilde{r}$, we have

$$D_N^{\text{MTP}}(\text{diag}(r) + \tilde{R}, 0) = D_N^{\text{MTP}}(\text{diag}(r_k + \tilde{r}), 0) = \sum_{k=1}^K \frac{\tilde{n}_k}{N} D_k(r_k + \tilde{r}_k)$$

(42)
where the first step is due to Lemma 4 and second step follows from Assumption 1. Combining this expression with the fact that \( \frac{1}{N}E[X^T X] = \text{diag}(n_1/N, \ldots, n_K/N) \), leads to the following simplification of approximation formula:

\[
\hat{D}_{\text{MTP}}^N(\text{diag}(r), \text{diag}(\text{vec}(\Lambda))) = \max \inf \left\{ \sum_{k=1}^K \frac{n_k}{N} D_k(r + \hat{r}_k) + \frac{1}{2} q^T \Lambda q - \frac{1}{2} \hat{r}_q q^T q \right\}.
\]  

(43)

Here, the maximum is over \( K \)-dimensional vectors satisfying \( 0 \leq q_k \leq n_k/N \) and the infimum is over \( K \)-dimensional vectors with non-negative entries. Finally, taking the \( N \to \infty \) limit of this expression leads to the formula for the limit given in Theorem 1.

### A.3 Proof of Theorem 2

The proof of Theorem 2 follows from fact that changes in the relative entropy with respect to the parameters \( (r, \Lambda) \) can be related to the MMSE in estimating the underlying variables. As a consequence of the I-MMSE relationship \([22]\), it can be shown that \( D_N \) is differentiable on the interior of its domain with partial derivatives:

\[
\partial_r D_N(r, \Lambda) = \frac{1}{2N} \text{E} \left[ \|E[x_k | Y]\|^2_2 \right], \quad \partial_{\lambda_{k\ell}} D_N(r, \Lambda) = \frac{1}{2N^2} \text{E} \left[ \|E[x_k x_\ell^T | Y]\|^2_2 \right].
\]  

(44)

These derivatives can be defined uniquely on the boundaries \( r_k = 0 \) or \( \lambda_{k\ell} = 0 \) according to the right derivatives, which are exist and are finite due the finite fourth moment assumption \([23, \text{Proposition 7}]\). The connection with the MMSE follows from the relationship

\[
\text{E} \left[ \|x_k - E[x_k | Y]\|^2_2 \right] = \text{E} \left[ \|x_k\|^2_2 \right] - \text{E} \left[ \|E[x_k | Y]\|^2_2 \right]
\]  

(45)

\[
\text{E} \left[ \|x_k x_\ell - E[x_k x_\ell^T | Y]\|^2_2 \right] = \text{E} \left[ \|x_k x_\ell^T\|^2_2 \right] - \text{E} \left[ \|E[x_k x_\ell^T | Y]\|^2_2 \right].
\]  

(46)

The assumption that the entries are independent with second moment equal to one means that \( \text{E} \left[ \|x_k\|^2_2 \right] = n_k \) and \( \text{E} \left[ \|x_k x_\ell\|^2_2 \right] = n_k n_\ell \) for \( k \neq \ell \). For the diagonal blocks, we can write \( \text{E} \left[ \|x_k x_\ell^T\|^2_2 \right] = n_k \xi_k + n_k(n_k - 1) \) where \( \xi_k \) is the fourth moment of \( P_k \) and so \( n_k^{-2} \text{E} \left[ \|x_k x_\ell^T\|^2_2 \right] \) converges to one in the \( N \to \infty \) limit.

The next step is to relate the derivatives of \( D_N \) to the derivatives of its limit \( D \). This is possible due to the fact that \( D_N \) is convex, which implies that \( D \) is convex, and hence differentiable almost everywhere, and furthermore that the derivatives converge at every point where the limit is differentiable \([24, \text{Proposition 4.3.4}]\).

With these results in hand, the asymptotic MMSE can be analyzing by studying the limit \( D \). If for a given pair \( (r, \Lambda) \) the maximum in the definition of \( D \) is attained at a unique point \( q^* \) then \( D \) has partial derivatives with respect to the nonzero entries in \( (r, \Lambda) \). Specifically, for \( r_k > 0 \) and \( \lambda_{r_k} > 0 \) it can be verified that

\[
\partial_{r_k} D(r, \Lambda) = \frac{1}{2} q_k^*, \quad \partial_{\lambda_{r_k}} D(r, \Lambda) = \frac{1}{2} q_k^* q_k^*.
\]  

(47)

For \( r = 0 \) and \( \lambda_{r_k} = 0 \) these expressions still hold for the right derivatives and it can be shown (see \([15, \text{Section II-B}]\)) that they provide one-sided bounds of the form

\[
\limsup_{N \to \infty} \frac{1}{N} \text{E} \left[ \|E[x_k | Y]\|^2_2 \right] \leq q_k^*, \quad \limsup_{N \to \infty} \frac{1}{N^2} \text{E} \left[ \|E[x_k x_\ell^T | Y]\|^2_2 \right] \leq q_k^* q_k^*.
\]  

(48)

Noting that these are equalities whenever the upper bound is equal to zero and combining with the arguments above leads to the conditions in Theorem 2.
B AMP for heteroskedastic rank-one estimation

This section considers approximate inference methods for the heteroskedastic rank-one observation model given by

\[ y_{ij} = \sqrt{\frac{\lambda_{ij}}{N}} x_i x_j + w_{ij}, \quad i, j = 1, \ldots, n \]  

(49)

where \((x_i)\) are unknown random variables, \((w_{ij})\) are independent standard Gaussian noise terms, and \((\lambda_{ij})\) and \(N\) are known parameters. Furthermore, it is assumed that the variables \((x_i)\) are independently (but not identically) distributed with \(x_i \sim P_i\) for \(i = 1, \ldots, n\). Note that the groupwise model \(\Pi\) corresponds to the special case where the parameters \((\lambda_{ij})\) and the distributions are constant within groups.

**Symmetrization and prepossessing** To simplify the exposition of the algorithms it is convenient to apply a prepossessing step that re-scales and symmetrizes the observations. Specifically, we define the modified observations \((\tilde{y}_{ij})\) and parameters \((\tilde{\lambda}_{ij})\) according to

\[ \tilde{y}_{ij} = \begin{cases} \sqrt{\tilde{\lambda}_{ij}} y_{ij} + \sqrt{\lambda_{ji}} y_{ji} & i \neq j \\ \sqrt{\lambda_{ii}} y_{ii} & i = j \end{cases}, \quad \tilde{\lambda}_{ij} = \begin{cases} \lambda_{ij} + \lambda_{ji}, & i \neq j \\ \lambda_{ii} & i = j \end{cases} \]  

(50)

By the orthogonal invariance of the the standard Gaussian distribution it follows that \((\tilde{y}_{ij})\) are sufficient statistics for estimation of \((x_i)\). Furthermore, the distribution of \((\tilde{y}_{ij})\) is given by the symmetric model

\[ \tilde{y}_{ij} = \frac{1}{\sqrt{N}} \tilde{\lambda}_{ij} x_i x_j + \tilde{\lambda}_{ij}^{1/2} \tilde{w}_{ij} \]

(51)

where \(\tilde{w}_{ij} = \tilde{w}_{ji}\) with \((\tilde{w}_{ij} : 1 \leq i \leq j \leq n)\) independent standard Gaussian variables. Using this parameterization, the likelihood function is proportional to

\[ \prod_{1 \leq i < j \leq n} \exp \left\{ \frac{\tilde{y}_{ij} x_i x_j}{\sqrt{N}} - \frac{\tilde{\lambda}_{ij} x_i^2 x_j^2}{2N} \right\} \]  

(52)

**Belief propagation** We begin with a derivation of the belief propagation algorithm \[42\]. To simplify the derivation we assume that that each probability measure \(P_i\) had a density \(f_i\) with respect to the Lebesgue measure. In view of \(\Pi\), the conditional distribution of the variables \((x_i)\) given the observations \((\tilde{y}_{ij})\) has a density that is proportional to

\[ \prod_{i=1}^{n} \phi_i(x_i) \prod_{1 \leq i < j \leq n} \psi_{ij} (x_i, x_j), \]

(53)

where

\[ \phi_i(x_i) := f_i(x_i) \exp \left\{ \frac{\tilde{y}_{ii} x_i^2}{\sqrt{N}} - \frac{\tilde{\lambda}_{ii} x_i^4}{2N} \right\}, \quad \psi_{ij} (x_i, x_j) := \exp \left\{ \frac{\tilde{y}_{ij} x_i x_j}{\sqrt{N}} - \frac{\tilde{\lambda}_{ij} x_i^2 x_j^2}{2N} \right\} \]  

(54)

The belief propagation algorithm is described by a collection of messages \((\mu_{ij} : 1 \leq i, j \leq n, i \neq j)\) where each \(\mu_{ij}\) is a non-negative function that describes the `influence' of variable \(i\) on variable \(j\). Starting at a given initialization, these messages are determined recursively via the update rule

\[ \mu_{ij}(x_j) \leftarrow \int \phi_i(x_i) \psi_{ij} (x_i, x_j) \prod_{k \in [n] \setminus \{i, j\}} \mu_{ki}(x_i) \, dx_i, \]

(55)
where \([n] = \{1 \ldots n\}\). The approximation to the marginal of the \(i\)-th variable associated with given collection of messages is given by the probability density function:

\[
x_i \mapsto \frac{\phi_i(x_i) \prod_{j \in [n]\setminus i} \mu_{ji}(x_i)}{\int \phi_i(x'_i) \prod_{j \in [n]\setminus i} \mu_{ji}(x'_j) \, dx'_i}.
\] (56)

Note that this final approximation is invariant to rescaling of the messages, and thus at each stage in the algorithm the message \(\mu_{ij}\) can be rescaled by an arbitrary positive constant.

**Relaxed belief propagation** Next we consider a relaxed belief propagation algorithm that is obtained by an approximation of the update rule. The first step in this approximation is to replace the function \(\psi_{ij}(x_i, x_j)\) by its second order expansion with respect to the term \((x_i x_j)/\sqrt{N}\) evaluated at zero:

\[
\psi_{ij}(x_i, x_j) \approx 1 + \frac{\tilde{y}_{ij} x_i x_j}{\sqrt{N}} + \frac{1}{2} \frac{(\tilde{y}_{ij}^2 - \tilde{\lambda}_{ij}) x_i^2 x_j^2}{N}.
\] (57)

Plugging this approximation into (55) leads to the following integral

\[
\int \phi_i(x_i) \left(1 + \frac{\tilde{y}_{ij} x_i x_j}{\sqrt{N}} + \frac{1}{2} \frac{(\tilde{y}_{ij}^2 - \tilde{\lambda}_{ij}) x_i^2 x_j^2}{N}\right) \prod_{k \in [n]\setminus\{i,j\}} \mu_{ki}(x_i) \, dx_i
\] (58)

To characterize the solution to this integral, define

\[
m_{ij} = \int x_i \phi_i(x) \prod_{k \in [n]\setminus\{i,j\}} \mu_{ki}(x_i) \, dx_i
\] (59)

\[
v_{ij} = \int x_i^2 \phi_i(x) \prod_{k \in [n]\setminus\{i,j\}} \mu_{ki}(x_i) \, dx_i - \bar{x}_{ij}^2
\] (60)

to be the mean and variance, respectively, of the probability density function

\[
x_i \mapsto \frac{\phi_i(x_i) \prod_{k \in [n]\setminus\{i,j\}} \mu_{ki}(x_i)}{\int \phi_i(x') \prod_{k \in [n]\setminus\{i,j\}} \mu_{ki}(x'_i) \, dx'_i}.
\] (61)

This density is similar to the marginal approximation in (59) except that message from the \(j\)-th variable is also excluded. Using this notation, the integral in (58) is proportional to

\[
1 + \left(\frac{\tilde{y}_{ij} m_{ij}}{\sqrt{N}}\right) x_j + \frac{1}{2} \left(\frac{\tilde{y}_{ij}^2 - \tilde{\lambda}_{ij}}{N}\right) \left(m_{ij}^2 + v_{ij}\right) x_j^2.
\] (62)

Using the approximation \(1 + au + \frac{1}{2}(a^2 - b)u^2 \approx \exp(au - \frac{1}{2}bu^2)\) for \(u \approx 0\) and recalling that we can rescale the message \(\mu_{ij}\) by an arbitrary positive constant leads to the modified update rule

\[
\mu_{ij}(x_j) \leftarrow \exp \left\{ a_{ij} x_j - \frac{1}{2} b_{ij} x_j^2 \right\}
\] (63)

where

\[
a_{ij} = \frac{\tilde{y}_{ij} m_{ij}}{\sqrt{N}}, \quad b_{ij} = \frac{\tilde{\lambda}_{ij} m_{ij}^2 + (\tilde{\lambda}_{ij} - \tilde{y}_{ij}^2)v_{ij}}{N}.
\] (64)

According to the update rule (63), the probability density in (61) is defined by the (unnormalized) density

\[
\phi_i(x_i) \prod_{k \in [n]\setminus\{i,j\}} \exp \left\{ a_{ki} x_i - \frac{1}{2} b_{ki} x_i^2 \right\} = \phi_i(x_i) \exp \left\{ \sum_{k \in [n]\setminus\{i,j\}} a_{ki} x_i - \frac{1}{2} b_{ki} x_i^2 \right\}
\] (65)
Here, we see that the products of the messages are described compactly in terms of a linear combination of the \( a_{ki} \) and \( b_{ki} \) terms.

The relaxed belief propagation can be described as follows. For \( i = 1, \ldots, n \) define the function
\[
\eta_i(a, b) = \frac{\int x_i \exp \{ax_i - \frac{1}{2}b_{ij}^2 \} \phi_i(x_i) \, dx_i}{\int \exp \{ax_i - \frac{1}{2}b_{ij}^2 \} \phi_i(x_i) \, dx_i}
\]
(66)
and let \( \eta_i'(a, b) \) denote the partial derivative with respect to the first argument. Note that \( \eta_i(a, b) \) and \( \eta_i'(a, b) \) represent the mean and variance, respectively, of the (unnormalized) density \( x_i \mapsto \phi_i(x_i) \exp(ax_i - \frac{1}{2}b_{ij}^2) \).

Starting with initial values for \((m_{ij}, v_{ij})\) the updates are defined according to
\[
m_{ij} \leftarrow \eta_i \left( \sum_{k \in [n]\setminus\{i,j\}} a_{ki}, \sum_{k \in [n]\setminus\{i,j\}} b_{ki} \right), \quad (67a)
\]
\[
v_{ij} \leftarrow \eta_i' \left( \sum_{k \in [n]\setminus\{i,j\}} a_{ki}, \sum_{k \in [n]\setminus\{i,j\}} b_{ki} \right), \quad (67b)
\]
where the terms \((a_{ki}, b_{ki})\) are defined with respect to the current values of \((m_{ki}, v_{ki})\) according to (64).

**Approximate message passing** One of the limitations of the relaxed belief propagation algorithm in (67a) is that it requires keeping track of \( n^2 \) terms. AMP can be viewed as an approximation to the relaxed belief propagation algorithm that requires only \( n \) terms.

Before stating the AMP algorithm, it is useful to consider the version of relaxed belief propagation where all of the updates are made in parallel. Specially, starting with an initialization \((m_{ij}^0, v_{ij}^0)\), the updates at time \( t = 1, 2, \ldots \) are given by
\[
a_{ij}^t = \frac{\tilde{y}_{ij}}{N} m_{ij}^t \quad (68a)
\]
\[
b_{ij}^t = \frac{\tilde{\lambda}_{ij}}{N} (m_{ij}^t)^2 + \frac{\tilde{\lambda}_{ki} - \tilde{y}_{ij}}{N} v_{ij}^t \quad (68b)
\]
\[
m_{ij}^{t+1} = \eta_i \left( \sum_{k \in [n]\setminus\{i,j\}} a_{ki}^t, \sum_{k \in [n]\setminus\{i,j\}} b_{ki}^t \right) \quad (68c)
\]
\[
v_{ij}^{t+1} = \eta_i' \left( \sum_{k \in [n]\setminus\{i,j\}} a_{ki}^t, \sum_{k \in [n]\setminus\{i,j\}} b_{ki}^t \right) \quad (68d)
\]
In comparison, the corresponding version of AMP is given by
\[
a_i^t = \frac{1}{\sqrt{N}} \sum_{k \in [n]\setminus\{i\}} \tilde{y}_{ki} m_k^t - \frac{1}{N} m_i^{t-1} \sum_{k \in [n]\setminus\{i\}} \tilde{y}_{ki} v_k^t \tilde{y}_{ki} \quad (69a)
\]
\[
b_i^t = \frac{1}{N} \sum_{k \in [n]\setminus\{i\}} \left( \tilde{\lambda}_{ki} (m_k^t)^2 + \tilde{\lambda}_{ki}^2 - \tilde{y}_{ki}^2 \right) v_k^t \quad (69b)
\]
\[
m_i^{t+1} = \eta_i(a_i^t, b_i^t) \quad (69c)
\]
\[
v_i^{t+1} = \eta_i'(a_i^t, b_i^t) \quad (69d)
\]
The relationship between these algorithms can be derived heuristically based on a decomposition of the means of the form \( m_{ij}^t = m_i^t + \epsilon_{ij}^t \) where \( m_i^t \) does not depend on \( j \) and \( \epsilon_{ij}^t \) is a small fluctuation; see e.g., [10, Appendix A].
To express the AMP algorithm using vector notation, we define $m^t$ and $v^t$ to be $n \times 1$ vectors containing the means and variances, respectively, at iteration $t$. Also, we define $\tilde{Y}$ and $\tilde{\Lambda}$ to be $n \times n$ symmetric matrices with diagonal entries equal to zero and off-diagonal entries given by $\tilde{y}_{ij}$ and $\tilde{\lambda}_{ij}$, respectively. The AMP algorithm in (69) can be written compactly as

\begin{align}
\alpha^t &= \frac{1}{\sqrt{N}} \tilde{Y} m^t - \frac{1}{N} (\tilde{Y} v^t) \circ m^{t-1} \\
\beta^t &= \frac{1}{N} \tilde{\Lambda} (m^t)^{\circ 2} + \frac{1}{N} (\tilde{\Lambda} - \tilde{Y} v^t) v^t \\
m^{t+1} &= \eta (\alpha^t, \beta^t) \\
v^{t+1} &= \eta' (\alpha^t, \beta^t)
\end{align}

(70a, 70b, 70c, 70d)

where $\eta$ and $\eta'$ are obtained by stacking the functions $\eta_i$ and $\eta'_i$, respectively, for $i = 1, \ldots, n$.

**Remark 2.** In the AMP algorithm given in (70) The dependence on distribution of $x_i$ is encapsulated by the function $\eta_i(a, b)$ defined in (66). This means that the algorithm can be applied in the case where $x_i$ is drawn according to a probability measure $P_i$ provided that the functions

\begin{align}
(a, b) \mapsto \int x_i^p \exp \left\{ a x_i - \frac{1}{2} b x_i^2 \right\} \exp \left\{ \frac{\tilde{y}_{ii} x_i^2}{\sqrt{N}} - \frac{\tilde{\lambda}_{ii} x_i^4}{2N} \right\} dP_i(x_i), \quad p = 0, 1, 2
\end{align}

(71)

can be approximated numerically.

**Remark 3.** If the parameter $\lambda_{ii}$ is small relative to the global scale parameter $N$, then the data point $\tilde{y}_{ii}$ has negligible impact and the term $\phi_i(x_i)$ in (66) can be replaced by the density of $x_i$. This is precisely what happens in the usual rank-one estimation model where all the $\lambda_{ij}$ are identical. By contrast, the heteroskatistic model (49) allows for the possibility that the diagonal terms $\lambda_{ii}$ are of the same order as $N$ and in this case, the diagonal terms should not be discarded.

**AMP for groupwise model** The AMP presented in Section 3.3 is obtained by specializing the AMP algorithm in (70) to the to setting of the groupwise model. Following the discussion in Remark 3 the function $\eta_i(a, b)$ defined in (66) is replaced by the definition given in (25), which omits the dependence on the diagonal entries of the observations.

## C Numerical approximation of relative entropy and MMSE

Numerically approximating the relative entropy and MMSE requires solving a low-dimensional saddle point problem and generating expressions for the single-letter relative entropy and its first derivative. In this work, we only consider numerical approximation of the formulas for the two-group case, as in (4). However, the single-letter relative entropy and its first derivative will be the same in the $K$-group case and our strategy for estimating the global optimum of the saddle point problem would be a reasonable strategy for $K > 2$ as well.

### C.1 Estimation of the global optimum

In the two-group setup, we only have to perform an optimization over four scalar variables in order to obtain the relative entropy and MMSE. To further simplify the process, we can note that the envelope theorem (see (36)) gives us that we need only consider the stationary points of the objective. At first glance of (14), this is far from obvious. In practice, we use a multivariate root finding algorithm from the SciPy optimization package in order to find the locations where the gradient is equal to zero.
C.2 Approximation of the single-letter relative entropy

In order to deal with saddle point problem in (14), we need to obtain a formula for the single-letter relative entropy, defined in (5), and its first derivative. In this work, we consider standard Gaussian priors and discrete priors. In the case of standard Gaussian priors, we have

\[ D_k(\gamma) = \frac{1}{2}(\gamma - \log(1 + \gamma)) \] and

\[ \partial_\gamma D_k(\gamma) = \frac{\gamma}{2(1+\gamma)}. \]

We also consider discrete priors which have the following form

\[ P_X = \sum_{i=1}^{M} p_i \delta_{a_i}, \tag{72} \]

where \( \delta_{a_i} \) denotes a point mass function at \( a_i \) and \( p_i \) are probabilities for \( i = 1, \ldots, M \) and \( P_X \) denotes the probability measure. In order to obtain the single-letter relative entropy, we consider the random variable \( Y = \sqrt{\gamma}X + Z \), where \( X \sim P_X \) and \( Z \sim \mathcal{N}(0,1) \) are independent. The probability density function of \( Y \) is a Gaussian mixture, which is given by

\[ f_Y(y) = \sum_{i=1}^{M} p_i \mathcal{N}(y; \sqrt{\gamma}a_i, 1) \tag{73} \]

where \( \mathcal{N}(y; \mu, \sigma^2) \) denotes the Gaussian probability density function with mean \( \mu \) and variance \( \sigma^2 \) evaluated at \( y \). The single-letter relative entropy function and its derivative can be expressed as one-dimensional Gaussian integrals, which are approximated numerically.